



# Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review

Rock Yuren Pang  
University of Washington  
Seattle, Washington, USA  
ypang2@cs.washington.edu

Hope Schroeder  
MIT  
Cambridge, Massachusetts, USA  
hopes@mit.edu

Kynneddy Simone Smith  
Columbia University  
New York, New York, USA  
simone.kynneddy@columbia.edu

Solon Barocas  
Microsoft Research  
New York, New York, USA  
sbarocas@cornell.edu

Ziang Xiao  
Johns Hopkins University  
Baltimore, Maryland, USA  
ziang.xiao@jhu.edu

Emily Tseng  
Microsoft Research  
New York, New York, USA  
etseng42@gmail.com

Danielle Bragg  
Microsoft Research  
Cambridge, Massachusetts, USA  
danielle.bragg@microsoft.com

## Abstract

Large language models (LLMs) have been positioned to revolutionize HCI, by reshaping not only the interfaces, design patterns, and sociotechnical systems that we study, but also the research practices we use. To-date, however, there has been little understanding of LLMs' uptake in HCI. We address this gap via a systematic literature review of 153 CHI papers from 2020-24 that engage with LLMs. We taxonomize: (1) domains where LLMs are applied; (2) roles of LLMs in HCI projects; (3) contribution types; and (4) acknowledged limitations and risks. We find LLM work in 10 diverse domains, primarily via empirical and artifact contributions. Authors use LLMs in five distinct roles, including as research tools or simulated users. Still, authors often raise validity and reproducibility concerns, and overwhelmingly study closed models. We outline opportunities to improve HCI research with and on LLMs, and provide guiding questions for researchers to consider the validity and appropriateness of LLM-related work.

## CCS Concepts

• **Human-centered computing** → **HCI theory, concepts and models**; **Empirical studies in HCI**; • **General and reference** → **Surveys and overviews**.

## Keywords

Large language models, human-AI interaction, systematic literature review, HCI theory

## ACM Reference Format:

Rock Yuren Pang, Hope Schroeder, Kynneddy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a

Systematic Literature Review. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26–May 01, 2025, Yokohama, Japan. ACM, New York, NY, USA, 20 pages. <https://doi.org/10.1145/3706598.3713726>

## 1 Introduction

Large language models (LLMs) are poised to transform the landscape of Human-Computer Interaction (HCI) research. Already, researchers have been using LLMs across the HCI research pipeline, from ideation and system development to data analysis and paper-writing [76]. Past work has shown rapid growth in the raw count of LLM-focused paper preprints, especially in HCI topics [118]. The explosion of LLM-related research has also led to rising discourse in HCI on the opportunities and challenges of LLM usage, including interview and survey studies with researchers to understand their practices [76], and workshops [4, 131] and social media commentary [67] in which scholars debate how the field ought to respond. The surge in LLM-related papers and discussions indicates a growing need to support scholars in understanding the potential and pitfalls of LLMs in HCI.

Such inquiry is consequential not only for HCI, but also for the broader landscape of computing research. On one hand, scholars in natural language processing (NLP) and machine learning (ML) increasingly look to incorporate human evaluation in LLM architectures, via techniques like reinforcement learning for human feedback (RLHF) that draw upon HCI methodologies [61, 92, 100, 172]. On the other hand, researchers across various communities, such as science and technology studies (STS), computer-supported cooperative work (CSCW), and fairness, accountability, and transparency (FAccT) have called for reflection on the potential negative impacts [35, 59, 144, 146], including a rising chorus of scholars exploring the societal implications of LLM development and the need for responsible AI practices [2, 35, 163]. As various research communities increasingly pursue human-centered methods and questions, there emerges an urgent need for we as the HCI community to scrutinize our own field, and to develop standards for researchers using LLMs and asking HCI-oriented questions. By scrutinizing how HCI methodologies are shaping and being shaped



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '25, Yokohama, Japan*

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1394-1/25/04

<https://doi.org/10.1145/3706598.3713726>

by LLM development, we can take a first step towards ensuring that their influence aligns with scientific rigor and societal benefit.

To this end, we contribute a systematic literature review of the 153 LLM-related papers in the last five years of CHI proceedings (2020–2024). Our key research questions include: Where have LLMs been applied at CHI? How have researchers used LLMs in their papers? What contributions have LLM-related scholarship made to HCI? What concerns around LLMs do authors articulate? Our intended audience includes both HCI researchers exploring LLM integration in HCI work and LLM practitioners seeking to understand the current best practices around using LLMs to interact with humans in various domains. We identify that LLMs have been taken up in 10 diverse application domains, including Communication & Writing, Augmenting Capabilities, Education, Responsible Computing, Programming, Reliability & Validity of LLMs, Well-being & Health, Design, Accessibility & Aging, and Creativity. Applying Wobbrock and Kientz [168]’s framework of research contributions in HCI, we found that LLMs were overwhelmingly used for empirical and artifact contributions, with limited work in theoretical or methodological advances. To characterize how LLMs are being integrated into HCI projects, we identified five roles that LLMs are playing in research projects: LLMs as system engines, LLMs as research tools, LLMs as participants or users, LLMs as objects of study (e.g., through audits), and users’ perceptions of LLMs (e.g., through interview studies of LLM users’ experiences). We also identified 22 common limitations and risks that authors acknowledged, ranging from qualms around LLMs’ performance to concerns around research validity, resource constraints, and potential consequences. We found that authors often raise validity and reproducibility concerns around LLM research, despite overwhelmingly studying closed-source LLMs.

Overall, this work presents an in-depth investigation of the current landscape of how HCI applies and studies LLMs. Towards more rigorous research and responsible design with LLMs, we outline directions for future HCI research at the LLM frontier, and provide actionable recommendations to researchers and practitioners. In summary, we contribute:

- a systematic literature review of 153 LLM-related papers from CHI proceedings 2020–2024, resulting in 10 domains where LLMs have been applied, 5 roles that LLMs play in HCI projects, and 29 limitations described by authors;
- opportunities for HCI research to leverage LLMs, including under-researched application domains, contribution types, and methodological gaps;
- guiding questions for HCI researchers to consider the *validity* and *appropriateness* of a proposed LLM-related study;
- an open-source dataset of 153 sampled papers from CHI 2020–2024 with our qualitative codes and paper metadata, publicly available at <https://llm-ification.github.io>.

## 2 Related Work

### 2.1 Literature Reviews in HCI

HCI has a rich tradition of using systematic literature reviews to identify patterns, trends, and limitations of a research area [147]. Such reviews provide conceptual frameworks for shared understanding across the field. Many prior works qualitatively analyzed

their paper samples to surface high-level themes. For example, Mack et al. [112] examined 836 accessibility papers over 26 years, coding for common contribution types, communities of focus, and methods. Stefanidi et al. [147] annotated 189 HCI literature surveys 1982–2022 to explain current contributions and topics within HCI. Similarly, Dell and Kumar [33] manually reviewed 259 HCI4D publications to provide an overview of the space. Caine [15] synthesized standards for sample sizes at CHI by manually extracting data from each CHI2014 manuscript. Quantitative methods have also been employed to provide broader perspectives on HCI research trends. Liu et al. [108] used hierarchical clustering, strategic analysis, and network analysis to map the evolution of major themes in HCI. Cao et al. [16] analyzed patent citations to study the relationship between HCI research and practice.

**Our work** builds on this literature to understand LLMs’ impact on HCI. We chose a qualitative approach to provide a deep formative understanding of this rapidly evolving landscape and its impact, not only for HCI researchers, reviewers, and students, but also for researchers in different communities (e.g., AI/NLP) who may be interested in the current state of LLM-ification in HCI, as well as practitioners looking for research-grade guidance on this rapidly evolving space.

### 2.2 Literature Reviews of LLM Papers

Outside of HCI, many fields across computing and social science have used literature reviews to study LLMs’ impact on their areas, including reviews of the models, the technical foci, and the societal implications of LLMs. Many of these reviews survey technical advancements, e.g., Zhao et al. [185] survey methods for training and evaluating core models, Gao et al. [44] review the state-of-the-art in retrieval-augmented generation, and Guo et al. [51] review multi-agent approaches. Other efforts have studied the risks posed by LLMs: Weidinger et al. [165] taxonomized the harms possible, including discrimination, information hazards, malicious uses, and environmental and economic harms.

Research has also surveyed trends in how LLMs are being applied in specific disciplines. Movva et al. [118] collected and analyzed 16,979 LLM-related papers posted to arXiv from 2018 to 2023 to understand trends in LLM research topics. Notably, they found that *society-facing* and *HCI* topics are the two fastest-growing, further showing the urgency of our focus on how the HCI community considers LLM use and implications. Movva et al. [118] also found that industry publishes an outside fraction of top-cited research, but also that industry papers tend to be less open about their models, datasets, and methods. Similarly, Fan et al. [37] used BERTopic to identify patterns in LLM research 2017–2023. Researchers have additionally employed topic modeling to study LLM usage in fields such as medicine [7] and education [103]. A recent study shows that papers in behavioral and social science disproportionately favor closed models, despite the availability of powerful, more reproducible open alternatives [170].

**Our work** focuses on CHI papers, to explore *where* authors applied LLMs in HCI research, and *how* authors leveraged them to make *what* contributions. We extend Movva et al. [118]’s quantitative work with an in-depth qualitative analysis of the HCI literature. Our focus on the last five years of CHI papers provides a window

into the most recent and most leading-edge work in HCI, since CHI has long been the central and most prestigious venue in the area.

### 2.3 How LLMs Can and Should Change Research

There has been substantial debate across the scientific community on how much LLMs can and should transform research [12]. Many papers argue that LLMs are poised to be incorporated in all disciplines, but call for consideration of their limitations. For instance, Aubin Le Qu  r   et al. [4]’s CHI’24 workshop discussed opportunities and responsible integration of LLMs into data work. In computational social science, researchers found that LLMs achieved fair agreement levels with humans on labeling tasks [187]. Researchers have also considered whether LLMs can or should influence academic writing [78]. A survey of 950,965 papers found a significant increase in the use of LLMs in writing scientific papers, especially in Computer Science [98]. However, many argue that researchers should “*avoid overreliance on LLMs and to foster practices of responsible science* [12].”

**Our work** extends the discussion on how LLMs are changing and should change research by focusing on the CHI community. We identify the unique roles that LLMs play in HCI research, analyze common limitations reported by authors, and advocate for proactive consideration of these limitations to ensure research rigor.

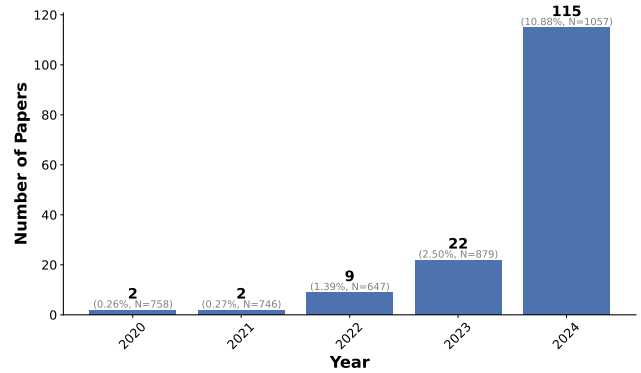
## 3 Methods

To understand the LLM-ification of CHI papers, we performed a literature review of CHI proceedings from 2020–2024. In our study, we focus on generative LLMs, rather than encoder-only models such as BERT or RoBERTa. Via iterative human coding, we assessed (1) the types of contributions common in LLM-focused HCI scholarship, (2) the roles that LLMs are playing in research projects; and (3) the limitations that researchers are disclosing in their papers.

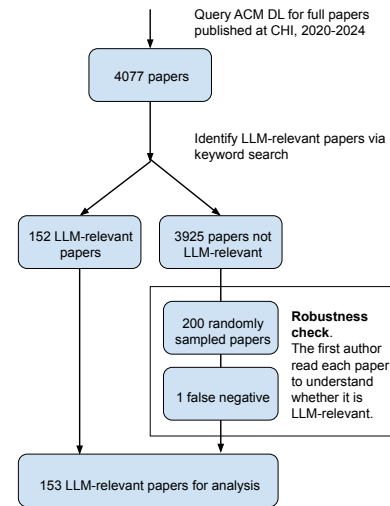
### 3.1 Data

We first gathered the full-text proceedings of CHI from 2020–2024, which at the time of writing represented the most recent five years of cutting-edge HCI research. In 2020, OpenAI’s GPT-3 was released, marking a leap in language models’ predictive capabilities. LLMs then became more accessible to researchers through APIs and open-sourced models.<sup>1</sup> We chose CHI for two reasons. First, CHI is the flagship international conference on HCI. All papers undergo rigorous peer review, and publications have significant impact on HCI research generally. Similar prior literature reviews chose CHI as a representative sample to identify trends in HCI [8, 104]. Second, CHI papers span a wide range of application areas and methodologies, (e.g., CHI 2024 had 16 subcommittees) giving this work broad representation. We acknowledge that ACM SIGCHI sponsors 26 HCI conferences,<sup>2</sup> which have more focused scopes. Our sample should be considered *generative* rather than exhaustive, and our work can spur future analysis of more focused conferences.

Our process follows an adapted PRISMA statement [117, 124] and is summarized in Figure 2. We began by assembling a corpus of CHI



**Figure 1: The raw number of LLM-related papers, followed by the percentage of the total number of papers in each year 2020–2024.**



**Figure 2: A flow diagram on our sample selection and refinement process.**

papers, and filtering for LLM-relevant works. We contacted ACM Digital Library (DL) in July 2024 for full papers (excluding extended abstracts, doctoral consortium submissions, and other non-paper artifacts) published at the conferences from 2020–24. This search resulted in an initial corpus of 4,077 papers. We then filtered each paper’s title and abstract by a set of keywords: “language model”, “llm”, “foundation model”, “foundational model”, “GPT”, “ChatGPT”, “Claude”, “Gemini”, “Falcon”. This filter resulted in a corpus of 152 LLM-relevant papers. Figure 1 shows the breakdown of papers, as well as the percentage of the paper numbers in each year. We did not search full text on the CHI proceedings because our early investigation showed that it resulted in substantially more false positives (e.g., a paper might have one sentence that mentions their implications “*in the age of LLMs*”). We also did not include general keywords (e.g., “artificial intelligence”), since our focus is papers that include LLMs, rather than capturing the

<sup>1</sup><https://techcrunch.com/2020/06/11/openai-makes-an-all-purpose-api-for-its-text-based-ai-capabilities/>

<sup>2</sup><https://sigchi.org/conferences/>

wider field of AI research, as has been studied in prior work on the human-AI interaction literature [2, 49, 175, 178].

We additionally ensured the robustness of our filtering procedure by validating our corpus against false negatives. We conducted a stratified sampling of 200 papers that were initially found to be *not* LLM-relevant. The first author then read each of the 200 papers to check whether the work was LLM-relevant. Our review found one paper ( $N=1$ , 0.5%) that was not identified by our keyword search procedure. This paper mentioned GPT-4 just once, in their method section, without mentioning any other keywords in our list. We added this paper to our final corpus ( $N=153$ , see Figure 2).

### 3.2 Analysis

To analyze the 153 papers, we applied an iterative process to develop a codebook. The initial codebook included deductive codes based on our four research questions. For two of our research questions, we used existing taxonomies to seed our codebooks: on the contribution type, we used Wobbrock and Kientz [168]’s taxonomy of research contributions in HCI, and on the application domains for each paper, we used a taxonomy from Stefanidi et al. [147]. For the rest of the questions—on the roles of the LLMs in each paper, the limitations and risks of the research—we generated initial codebooks during the iterative open coding process.

We conducted four iterations of independently applying and updating the codebook, using a randomly selected set of 10 papers for each iteration. After each set, the research team came together to refine or merge existing codes, add new codes, and resolve disagreement through consensus. Throughout, we computed interrater reliability (IRR) using Krippendorff’s alpha to guide our discussions.<sup>3</sup> The final alpha values are  $\alpha_{\text{Contribution Types}} = 0.866$ ,  $\alpha_{\text{Application Domains}} = 0.849$ ,  $\alpha_{\text{LLM roles}} = 0.773$ ,  $\alpha_{\text{Limitations \& Risks}} = 0.887$ <sup>4</sup>. All values are comparable to prior work [90, 112]. This process led to a codebook of 51 low-level codes (see Supplementary Materials for the process and the codebook). Finally, the remaining papers—those that had not been used for codebook development—were split into three sets and coded independently by the three researchers who all participated in the codebook development. During this step, the authors met regularly to discuss any emergent concerns, and disagreements were resolved through consensus.

### 3.3 Research Positionality

We acknowledge that our academic and professional backgrounds have shaped our perspectives on this topic. All authors had experience using LLMs directly or studying users’ perceptions of LLM-powered systems, and had experience working in responsible computing. The authors’ expertise covers fields including HCI, NLP, computational social science, accessibility, machine learning, fairness, sociotechnical systems, and usable security and privacy. Collectively, we are US-based researchers at three different R1 universities and one US-based research institute.

<sup>3</sup>Note that Fleiss’ kappa can also be applied to the IRR analysis of the complete nominal data in our case. We chose Krippendorff’s alpha in line with prior literature review [112].

<sup>4</sup>The  $\alpha$  over our initial 29 low-level code is 0.633.

## 4 Results

Our analysis reveals *where* LLMs have been applied at CHI, *how* researchers have leveraged these models, and *what* contributions they made to the field of HCI. In parallel, we taxonomize the common *limitations and risks* articulated by authors (see Table 1).

### 4.1 Application Domains

We found 10 diverse domains in which HCI researchers have applied LLMs (Table 1). We elaborate each in this section.

**Communication & Writing** (22.88%,  $N=35$ ): This domain emerges as the most-studied area, spanning both specific writing tasks and *AI-mediated communication (AIMC)* [56], in which intelligent agents modify, augment, or generate messages to achieve communication goals. Many of these works imagine writers as the target LLM user, in tasks from personal diaries [80] and email composition [14] to storytelling [25], screenplay creation [116], and general creative writing [18, 161]. For instance, researchers have examined writers’ attitudes toward collaborating with LLMs [96], including how writers choose prompting strategies [31] and users’ perception of AIMC support in a variety of writing tasks, such as idea generation, translation, and proofreading [42]. Researchers have also examined how LLMs might introduce implicit bias to the writing process [41, 69].

**Augmenting Capabilities** (16.99%,  $N=26$ ): This domain includes papers that develop technologies to enhance human *performance* and *productivity* by altering how we engage with technology and information. Some attempt to bridge the physical and digital worlds in scenarios such as video conferencing [106] and mixed reality [32]. Many also study the future of work and productivity. Fok et al. [40] leverages LLMs to support sensemaking on business document collections, while Kobiella et al. [84] studied how ChatGPT usage affects professionals’ perceptions of workplace productivity. Several papers also developed tools to enhance productivity in academic research, building new approaches for sensemaking of literature [93] and research idea generation [160].

**Education** (14.38%,  $N=22$ ): This domain explores the potential of LLMs to enhance learning experiences for students and improve pedagogical methods for educators. For students, research examined learners’ existing interactions with LLMs, including Belghith et al. [9]’s investigation of middle schoolers’ approaches to and conceptions of ChatGPT. Several works explored using LLMs as learning aids in specific subject areas, such as math [181], vocabulary acquisition [89], and programming [72]. For educators, studies examined the LLMs’ integration into teaching. Han et al. [54] found that teachers are excited about potential benefits, namely LLMs’ ability to generate teaching materials and provide personalized feedback to students; however, teachers and parents are both concerned about their impact on students’ agency in learning, and potential exposure to bias and misinformation. Researchers have also designed LLM-based tools to assist teachers in domains such as cyberbullying education [60] and environmental science instruction [22].

**Responsible Computing** (12.42%,  $N=19$ ): This explores ethical and societal implications of computing systems, particularly in high-stakes domains and for vulnerable populations. It touches on issues like fairness, information hazards, and privacy. Several studies have examined how marginalized groups perceive LLMs, focusing on



	Code	Definition
	<i>Where have LLMs been applied in CHI papers?</i>	
Application Domains	<b>Communication &amp; Writing</b>	On various writing and communication tasks, which often target writers as the primary user groups.
	<b>Augmenting Capabilities</b>	On technologies to enhance human performance and productivity, often in the physical world.
	<b>Education</b>	On learning experiences for students and pedagogical methods for educators.
	<b>Responsible Computing</b>	On ethical and societal implications of computational systems, particularly in high-stakes domains.
	<b>Programming</b>	On various aspects of software development and programming tasks.
	<b>Reliability &amp; Validity of LLMs</b>	On evaluating and improving LLM outputs themselves.
	<b>Well-being &amp; Health</b>	On managing health-related disorders/illnesses, or interactions with health data or healthcare providers.
	<b>Design</b>	On various types of design work, which often target designers as the primary user group.
	<b>Accessibility &amp; Aging</b>	On population with disabilities and older adults.
	<b>Creativity</b>	On the creativity process and creativity support tools, which often overlaps with other domains.
	<i>How do CHI papers leverage these models?</i>	
LLM Roles	<b>LLMs as system engines</b>	LLMs function as core elements within systems, prototypes, algorithms, and programming frameworks.
	<b>LLMs as research tools</b>	LLMs perform research tasks traditionally executed by researchers in a research project, such as data collection, analysis, and writing.
	<b>LLMs as participants &amp; users</b>	LLMs simulate human responses and behaviors, or act as users or participants in an interaction.
	<b>LLMs as objects of study</b>	LLMs' inner mechanism, properties, performance are evaluated.
	<b>Users' perceptions of LLMs</b>	LLMs or tools (e.g., ChatGPT) are studied to understand user perceptions in different contexts.
	<i>What are the concerns by the authors at CHI?</i>	
Limitations & Risks	<b>Limitations on LLM Performance</b>	Limitations specifically on the LLM capability to output the desired output. This includes <i>LLM bias toward different groups, limited data coverage in the training data, non-deterministic response, hallucination, unspecific errors and biases</i> .
	<b>Limitations on Research Validity</b>	Limitations to the extent which an instrument measures what it claims to measure in the paper. This includes internal and/or external validity across users, contexts, models, and prompts.
	<b>Limitations on Resource</b>	Limitations on computational and financial resources to open or closed source LLMs. This includes <i>computational cost, financial cost, lack of evaluation standards</i> .
	<b>Risks to Society</b>	Potential negative and long-term outcomes, risks, or unintended effects may arise from the artifact or study. This includes <i>economic harms, representational harms, misinformation harms, malicious use, hate speech, and environmental harms</i> .

**Table 1: Domains where LLM applications are developed, roles of LLMs in HCI projects, and acknowledged risks and limitations. Note that we did not include contribution types in this table. A paper can have *multiple* (sub-)codes.**

gender [111, 150], religion [130], and other intersectionalities [43]. Research also identified the risks LLMs pose to those seeking information online. For instance, Sharma et al. [143] investigated how LLM-powered search systems might amplify echo chambers, while Oak and Shafiq [121] studied the use of LLMs by underground incentivized review services. Zhou et al. [186] outlined approaches to addressing LLM-generated misinformation. Papers also addressed a range of privacy issues, including online surveillance on social media [24], users' navigation of disclosure risks and benefits when using LLM-based conversational agents [184], and general privacy knowledge [20]. Finally, we identified papers that integrate LLMs into interactive tools designed to facilitate responsible computing practices [126, 163].

**Programming** (11.11%, N=17): This domain automates and improves software development and programming tasks, including papers related to data science, analytics, and visualization systems. Many papers develop tools to facilitate code creation. For instance, Liu et al. [105] introduced a novel method called grounded abstraction matching, powered by Codex, to assist non-expert programmers in guiding code generation. Other tools support programmers by providing no-code platforms for traditionally complex programming languages [81], explaining code generation [174], and aiding in programming language learning [21]. We also include work on “prompt engineering” in this category, such as prompt sharing [38], direct manipulation of LLM outputs [114], and visual prompt comparison [3]. These studies used programming tasks for evaluation,

reflecting the broader trend of incorporating prompt engineering into software engineering [167]. On a critical note, Kabir et al. [75] analyzed ChatGPT's responses to 517 StackOverflow programming questions, revealing that 52% of the answers contained incorrect information and 77% were verbose.

**Reliability & Validity of LLMs** (10.46%, N=16): This domain focuses on evaluating and improving LLM outputs. The first stream of work includes analyses determining the validity of applying LLMs to specific contexts. For example, He et al. [58] compared GPT-4 and Mechanical Turk pipelines for sentence labeling tasks from scholarly articles, showing that combining crowd and GPT-4 labeling increases accuracy. Another example is Kabir et al. [75], evaluating the validity of using LLMs' to answer programming questions. The second stream involves tools designed to enhance the reliability or validity of LLM outputs. For instance, HILL identifies and highlights hallucinations in LLM responses, allowing users to handle responses with greater caution [94]. EvalLM enables interactive evaluation of LLM outputs based on user-defined criteria across multiple prompts [82]. AI Chain is a visual programming tool for crafting LLM prompts, which improved the quality of task outcomes as well as the transparency, controllability, and the sense of collaboration when interacting with the black-box LLMs [169]. Zamfirescu-Pereira et al. [180] built a design tool to study how non-experts intuitively approached and evaluated prompt design.

**Well-being & Health** (9.15%, N=14): This domain refers to the management and prevention of health-related disorders and illnesses, or interactions with health data or with healthcare providers.<sup>5</sup> One thread of work involves assisting practitioners in providing better care. For example, Yang et al. [177] designed a GPT-3-based decision support tool that draws on the biomedical literature to generate AI suggestions. Yildirim et al. [179] worked with radiologists to explore the design space for incorporating LLMs into radiology. Another thread involves support for patients in self-tracking, self-diagnosing, and self-managing their illnesses. For instance, Sharma et al. [142] used a fine-tuned GPT-3 model to improve self-guided mental health interventions through cognitive restructuring, a technique to overcome negative thinking. MindfulDiary leveraged GPT-4 to support psychiatric patients' journaling [79]. Strömel et al. [148] found that GPT-generated data description can effectively complement numeric fitness data.

**Design** (8.50%, N=13): This domain captures papers whose target audience is designers. For example, HCI researchers have produced LLM-powered tools that facilitate the design process for practitioners, such as mobile UI design [36, 65, 171], landscape design [66], interior color design [63], and multimodal application design [101]. On the other hand, Liao et al. [99] interviewed 23 UX practitioners to explore the design space around LLMs supporting ideation, including their needs around model transparency.

**Accessibility & Aging** (7.84%, N=12): This domain focuses on people with disabilities and older adults. We found diverse accessibility contexts, including the blind or low-vision (BLV) community [152, 182], people with autism [70], learning disabilities involving Augmentative and Alternative Communication (AAC) [157], and situational impairments [107], as well as papers on older adults [173]. However, we did not find papers on the deaf or hard of hearing (DHH) community or motor or physical impairments, which are generally the second and third most prevalent in terms of accessibility paper counts [112].

**Creativity** (5.88%, N=9): This domain covers the creative process and creativity support tools. Chakrabarty et al. [18] proposed the Torrance Test of Creative Writing (TTCW) to directly scrutinize whether LLMs are "creative" via a story writing task. Similarly, Jigsaw presented a creativity support tool to assist *designers* with prototyping multimodal applications by chaining multiple generative models [101].

## 4.2 Contribution Types

The above application domains were primarily addressed through (1) empirical contributions (98.70%, N=151)—often to understand a population's view toward or use of LLMs or specific LLM-powered tools—and (2) artifact contributions (61.44%, N=94), which involve building a tool. These two contribution types frequently occur in combination, in studies where authors first build an artifact and then empirically test it with users. For artifact contributions, we observed that LLM-powered systems have a wide range of fidelity levels, from fully open-sourced systems with GitHub repositories to simple wireframes. The dominance of LLMs in these systems also varied, with some systems using LLMs throughout the entire pipeline and others using them only for processing textual data. We applied the code "*artifact contribution*" to a paper when authors claimed that LLMs are (or would be) a part of the system. The high frequency of artifact contribution (61.44% in our sample in contrast to 24.50% at CHI overall [168]) may indicate that LLMs might have lowered the barrier to prototype research artifact of high quality, a point we unpack further in 5.1.3.

The remaining five contribution types occurred less frequently, with one survey contribution and no opinion contributions. Distinguishing between methodological and artifact contributions can be challenging, as some methods are embedded in a system. Per [168], we used methodological contribution to refer to research method contributions in HCI. Methods for creating multimodal mobile applications, for example, were not included. Overall, we found 16 (10.46%) methodological contributions, such as LLM-augmented methods to enhance UX evaluation [85], generate synthetic user data [160], and provide metrics to measure creativity in LLMs [18]. We found 8 theoretical contributions (5.23%), ranging from a framework for collaborative group-AI brain-writing [140], a conceptual framework to bridge the gulf of envisioning [149], and a design space for intelligent writing assistants [90] (also a metareview contribution). Dataset contributions were less common (N=6, 4.0%). In the **LLM roles** section, several papers used LLMs to generate synthetic datasets, which may lower the barrier to creating large, diverse datasets for thorough evaluation, yet curating benchmark

<sup>5</sup>While some health-related conditions may fall under accessibility, such as chronic illness [47], we decide according to how the condition was treated: papers that adopt a social model of the condition or disability (i.e. that the incompatible design of society with the person's condition is the "problem") are *Accessibility*, and those that adopt a medical model (i.e. that the person's condition is the "problem") are classified here under *Well-being & Health* [52].

datasets of real users that can test the performance of LLMs at scale remains a challenge [91].

### 4.3 LLM Roles

We identified five roles that LLMs play in HCI research (Figure 3). While Figure 3 may not fit every project given the interdisciplinary nature of our field, it reflects our sample, which primarily offers empirical contributions.

**LLMs as system engines** (62.74%, N=96): In this role, LLMs function as core elements within systems, prototypes, algorithms, and programming frameworks. One way LLMs can be used in systems is to *generate content*, e.g., ideas, code, and conversations. For example, Farsight used LLMs to generate ideas to identify potential harms of AI applications [163], and GenLine used LaMDA to generate code from users' natural language [71]. MindTalker, a GPT-4 conversational agent, supports people with early-stage dementia by reducing loneliness [173]. On the other hand, LLMs may be used to *process information and extract insights*, e.g., by retrieving or summarizing from large, unstructured datasets. For instance, PaperWeaver deduces users' research interests from their paper collection on Semantic Scholar [93], while Memoro interprets user needs from the users' conversation history [188]. Visual Captions employed a fine-tuned LLM to predict user intent using the sentences in a video conferencing call [106]. Systems integrate LLMs at different levels. Some systems' main functions rely on a carefully-designed system prompt, often in a form instructions to a conversational agent [154], while others used LLMs as one [57] or more [87] step(s) in a complex pipeline. On another axis, the LLM-powered tools can range from a fully-functioned open-source system [163] to design prototypes that elicit important empirical insights [177].

**LLMs as research tools** (9.15%, N=15): We found several authors used LLMs to perform tasks traditionally executed by researchers or research assistants, including *data collection, analysis, or writing*. For example, Choksi et al. [24] applied LLMs to conduct qualitative coding on social media posts on NextDoor. They first developed a codebook, manually labeled 340 posts, and then adjusted the codebook prompts before using GPT-4 to tag the rest of the posts from the sample. Such LLM-augmented workflows were often also claimed as a methodological contribution, or packaged as a system that other researchers could use. For example, Wang et al. [162] introduced a multi-step human-LLM collaborative method for qualitative coding. In this process, LLMs generate labels and explanations, a verifier model assesses the quality, and humans re-annotate the subset of low-quality labels. Ding et al. [34] contributed a LLM-based method to identify critical online discussion patterns at scale to inform national health outcomes. Similarly, Lam et al. [87] proposed LLoM, a LLM-powered Python package to iteratively synthesize concepts over a sample of text.

Another thread involved using LLMs to *generate data* for research purposes. For example, Sun et al. [150] used GPT-2 to generate a corpus of 96,600 artificial greeting messages to study gender bias in greeting card messages and facilitate future research on this topic. Ko et al. [83] introduced a LLM-based framework that takes Vega-Lite specification as input to generate diverse natural language datasets, such as captions, utterances, and questions about the visualization. Feng et al. [39] uses LLMs to automatically mine UI

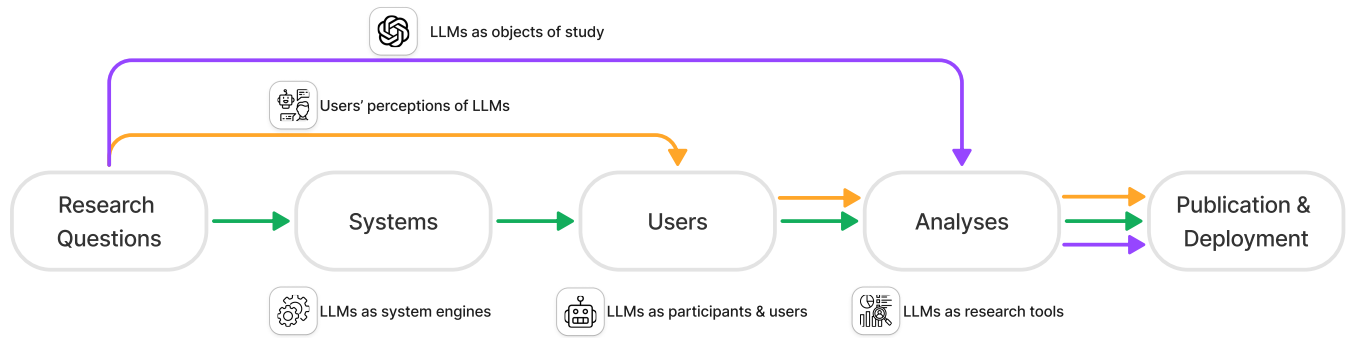
data from Android apps. These papers often generate synthetic datasets and conduct analyses as part of their contributions.

**Additional Analysis:** As using LLMs to perform research tasks is becoming new research methodology [4], we examined authors' justification of this role. For all 15 papers, authors justified their LLM usage, explained LLMs' capabilities and suitability for the task, and cited relevant prior work. All but one paper provided further experimental validation. These validations took the form of comparison user studies with non-LLM baselines, manual validation of system outputs, and human or computational quality assessments. All but one paper relied on humans for the evaluation, and this exception [150] used computational methods for quality analysis.

**LLMs as participants & users** (7.19%, N=11): This category uses LLMs to simulate human responses and behaviors, acting as users or participants. One line of work relied on the assumption that LLMs can create believable proxies for human behaviors, known as "*personas*". Personas were proposed decades ago in HCI [132] to guide user research, by creating abstract representations of users that provide valuable insights into user needs, behaviors, and preferences. By prompting LLMs to create such personas, researchers aim to approximate user feedback. For example, Impressona generated on-demand feedback from writer-defined AI personas of their target audience [11]. Similarly, Hedderich et al. [60] built Co-Pilot for teachers to prepare them to chat with students about cyberbullying. Another thread includes works on using LLMs to simulate potential user feedback for systems or designs. Duan et al. [36] applied GPT-4 to automate heuristic evaluation via a Figma plugin. Likewise, SimUser leveraged LLMs to simulate usability feedback [171]. Hämäläinen et al. [53] explicitly studied the validity of using LLMs to generate synthetic user research data, concluding that GPT-3 can generate believable answers to open-ended questionnaires about experiencing video games as art.

**Additional Analysis:** Using LLMs as participants & users constitutes a novel research methodology, so we also analyzed authors' justification of their methodology. We found that 10/11 papers provided both textual justification and experimental validation. Similar to the LLM-as-research-tools papers, the text justifications are also supported by citations to relevant prior work in NLP. Experiments similarly spanned user studies, as well as human and computational analysis. Rather than justifying the usage, Cuadra et al. [29] studied this very topic with a more critical lens and demonstrated the validity concerns inherent to LLM use in chatbots as humans, which Wang et al. [158] and Agnew et al. [1] address from an ethical perspective as well.

**LLMs as objects of study** (9.80%, N=15): This category contains papers that explore LLMs' underlying mechanisms and properties, including training datasets, response outputs, and issues (e.g., hallucination). Some works study potential problems inherent to LLMs themselves. For instance, Precel et al. [130] scrutinized common LLM training datasets and found that a disproportionate amount of content authored by Jewish Americans is used for training without their consent, and Sun et al. [150] studied the gender bias of GPT-2 generated text. Other works focus on the ecological validity of applying LLMs in a particular context. Kabir et al. [75] conducted an empirical study of the characteristics of ChatGPT answers to StackOverflow questions, evaluating whether LLMs



**Figure 3: Overview of roles that LLMs can play throughout common HCI research stages. LLM roles are depicted with icons and text, and arrows represent common empirical studies: → system building studies; → user studies; → data science studies.**

are appropriate to use in the context of Q&A for programming questions. Several studies examine the validity of using LLMs for crowdsourcing tasks [58].

**Users' Perceptions of LLMs** (23.53%, 36): This category includes studies on how users perceive LLMs or LLM-powered tools. Papers often examine a particular population's perception and usage of a public LLM-powered tools (e.g., ChatGPT) to create a design space or surface challenges and opportunities. For fine-grained insights, we exclude user studies that evaluate a system where the research artifact is the main contribution. For example, papers have studied how LGBTQ+ individuals experience using chatbots for mental health support [111]. Using the case of CareCall—a deployed chatbot for socially isolated individuals in South Korea—Jo et al. [73] attempted to understand how LLM-driven chatbots can support public interventions. Other works have studied how diverse users perceive and interact with LLMs or LLM-powered chatbots, including teachers [60, 153], middle schoolers [9], creative writers [45], and performance artists [74]. Several works have also examined LLMs' effects on users. For example, Wester et al. [166] studied how LLMs deny user requests, and Jakesch et al. [69] examined how users write social media posts with LLM assistance.

#### 4.4 Limitations

This section covers four top-level codes and 22 main sub-level codes for the limitations and risks discussed in our sample. Coding the limitations is not a trivial task, as not every paper has a dedicated "limitations" section. We found 94.77% (N=145) papers with a dedicated section for limitations (i.e., with "limitations" in the section title) and 14.38% (N=22) papers with a dedicated ethics or impact statement. Our analysis was primarily based on the limitations section; if there was no limitations section, we read through the paper to find potential mentions of limitations.

**4.4.1 LLM Performance** (42.48%, N=65). The top-level code refers to limitations on LLMs' capability to generate the desired output. These limitations highlight areas where the LLM's performance may fall short of expectations.

**LLM bias toward different groups** (11.11%, N=17): This limitation recognizes that LLMs' disparate representation across different populations. For example, Shin et al. [145] noted the GPT-3

and DALLÉ-2 in their system might output and perpetuate gender and racial stereotypes, including a higher chance of featuring white men rather than users in other racial groups. This limitation also includes cases where LLMs *fail* to model certain user groups—the absence of those users. Ma et al. [111] stated that LLM-based chatbots failed to "*recognize complex and nuanced LGBTQ+ identities and experiences, rendering the chatbots' suggestions generic and emotionally disengaged.*"

**Limited data coverage in the training data** (9.80%, N=15): Authors explicitly mentioned that LLMs' training data might be insufficient or outdated. For instance, Lee et al. [89] found they needed extra engineering steps to use an LLM with their Korean-speaking participants, which they attributed to "*GPT-4's underperformance in non-English languages*". When prompting LLM conversational agents to display empathy using elicitations from Reddit, Cuadra et al. [29] acknowledged that they are not aware of the distribution of the training data, and are therefore unable to tell whether the data used in the study has been covered by GPT-4.

**Non-deterministic response** (7.84%, N=12): Authors often recognized that LLM responses are probabilistic, and could change unpredictably even when given the same prompt. Gu et al. [50] recognized that their LLM's explanations were not fully controlled, because they used real-time responses from commercial models. Chen et al. [20] attributed the inconsistency of generated data to the "*inherent randomness embedded in the output of LLMs*." This, however, can be alleviated by changing the sampling temperature to zero [123] or using guided generation [97].

**Hallucination** (8.50%, N=13): LLMs can produce inaccurate or entirely fabricated information. Hoque et al. [62] explicitly pointed out that "*LLMs can generate hallucinations,*" which may "*alter the dynamic for such authors [in their study] when using an LLMs*" but later stated that studying the effect is out of their study scope. Though Retrieval Augmented Generation (RAG) systems may help alleviate this problem in the future [44, 68], applications that leverage this approach can still suffer from hallucination issues. For instance, Zulfikar et al. [188] stated that using LLMs "*in information retrieval can lead to hallucinated answers that do not exist in the dataset.*" To ensure validity, works such as PaperWeaver [93] attempted to evaluate the system's performance against hallucination by collecting



annotations of factual correctness for 60 descriptions in their study, but not all papers grappled with this problem as explicitly.

**Unspecified errors and biases** (16.99%, N=26): This the most common code related to LLM performance. Papers vaguely recognize the problems of LLMs, but authors did not specify the exact errors due to the models' black-box nature. For example, Li et al. [95] stated that *"like many other AI-based predictions, our system makes errors"* after users reported that the system's prediction did not match the intention. However, the author did not explain the potential errors caused by LLMs. Ko et al. [83] mentioned that *"the opaque nature of these models implies that we cannot have full control over their outputs or ensure exact replication in future studies."* Often, authors observed abnormal and inaccurate output from the systems and speculated the reasons for such underperformance. For example, Wang et al. [159] explained the underperformance is that *"LLMs are trained to generate text instead of the domain-specific task (i.e., selecting an element id in mobile UI)." Papers often defer addressing these errors to the future. Buschek et al. [14] acknowledged that their work is still a prototype, and suggested quality could be further improved through finetuning or training, "possibly involving even larger (email) datasets, extensive architecture search, or generally scaling up."*

**4.4.2 Resource Limitation** (28.76%, N=44). This top-level code refers to computational and financial resources needed to run LLMs, as well as a lack of evaluation standard or metrics. High resource demands can impact the efficiency and scalability of deploying the LLM, and can affect our community's ability to consistently evaluate LLMs or tackle common problems such as hallucination.

**Computational cost** (9.15%, N=14): Computational cost refers to the computational resources required to run LLMs, including the need for hardware (e.g., GPUs) for local execution and limited token windows, which restrict the amount of possible input. For example, Nguyen et al. [120], who employed OpenAI's Codex, wished that they had used open-source models to ensure study reproducibility, but recognized that doing so would *"impose significant computational requirements"* due to the need for extensive GPU resources. When facing the limited token size, authors had to devise workarounds. For example, Petridis et al. [128] split their documents into sections to accommodate GPT-3's input length, and wrote that that might have *"affected the overall performance and user experience of the system."*

**Financial cost** (3.27%, N=5): This resource constraint included monetary expenses with using LLMs, often tied to API calls for closed-source models and using online platforms like ChatGPT. For example, RELIC integrated GPT-3 due to its high performance, but authors also recognized that the LLM-enhanced component via the API *"will inevitably increase calculation expenses."* [23] Similarly, financial cost also impacts access to advanced chatbot playgrounds. In a study of ChatGPT's ability to answer programming questions, Kabir et al. [75] noted the \$20 per month subscription fee is a *"considerably high monetary value for many countries,"* and decided to use the free version (GPT-3.5) to lower the barrier for reproducibility at the expense of potential performance.

**Lack of evaluation standards/metrics** (16.99%, N=26): This category includes authors wishing to evaluate LLM aspects, but lacking the appropriate standard or metrics. A paper falls under this

category only when authors explicitly called for more standards (e.g., "open question" or "active research area"). For instance, Taeb et al. [152] recognized that some participants in their user study spotted errors in their system, but stated that *"evaluating the correctness of LLM-based systems remain an active area of research."* Cheng et al. [22] mentioned that guardrail the *safety* of their LLM-powered tool "without supervision" in the wild is still an "active research area". In the same paper, Cheng et al. [22] recognized that achieving ideal conversational context was still challenging, *"despite the abundance of literature on effectively engineering prompt for LLMs."* Several papers also called out a lack of benchmarks for evaluating LLMs outputs, such as conducting thematic analysis [87] and in mental health applications [79].

**4.4.3 Research Validity** (90.85%, N=139). Research validity is often defined as the extent to which an instrument measures what it claims to measure or if the study design can effectively test their hypotheses [113]. *Internal validity* refers to the legitimacy of a study's results, considering factors such as group selection, data recording methods, and analysis procedures [113]. *External validity* concerns the findings' transferability to other contexts of interest [113]. We consider ecological validity a subset of external validity, in that it refers to whether the studies resemble "real-world" conditions [138]. Validity issues can arise across users, contexts, LLMs, and prompts. In total, we identified  $2 \times 4$  codes related to this limitation. During coding, we first determined whether the issue impacted internal or external validity, and then identified the affected dimensions. We avoided assessing whether the project could have validity issues, but instead coded what the authors acknowledged in their paper.

The most prevalent limitations are **internal and external validity across users and contexts**. Internal validity issues related to users often stemmed from limited sample sizes and lack of diversity within samples. For example, Lin et al. [102] mentioned that *"a relatively small sample size leads to challenges in concluding some of the potential correlation."* This, in turn, may have external validity concerns. For example, Park and Ahn [127] mentioned that their research is based only on English-speaking university students, so the result *"may not reflect students who speak English as a second language."* Similar issues can also apply to different contexts, such as application scenarios. Zhang et al. [181] recognized that their study setup might have *"constrained the natural spontaneity that a human can bring to the storytelling process,"* which may have hurt the internal validity of observing behaviors that the authors claimed to study. Zhang et al. [182] acknowledged that their insights *"may or may not generalize to use in the field,"* because their prototype design constrained *"what tasks our participants could do."* Research validity issues across users and contexts are generally related to study designs evaluating LLMs or LLM-powered systems.

*Of the 153 papers, 130 papers (84.98%) used or studied a variation of the closed GPT-family models.* Despite this, many researchers articulated the **research validity issues across models**. Internal validity issues may arise when using LLMs. For example, Chakrabarty et al. [18] employed the default GPT-4 generation parameters (i.e., temperature = 1) to evaluate the model's capabilities. However, they recognized that a variation in temperature might have changed the content quality, thus affecting the study conclusion. Dang et al. [31] also acknowledged that they might not have identified the best

settings for model usage due to using the black-box models, which may affect the results internally.

The majority of the research validity issues are around external validity. For example, Dang et al. [31] addresses external validity in the same paper, stating that there might be “*potential changes to the model over time*”, which limit the exact replicability for their studies “*beyond our control*”. Kobiella et al. [84] conducted their study with GPT-3.5 and recognized that “*some findings might not be as prevalent*” with the release of GPT-4.

Given this external validity concern, many papers designed their system to be “*modular*” on purpose — swapping the underlying model with other models or even future, non-existing models. For instance, Feng et al. [39] mentioned that “*while we use the gpt-3.5-turbo as our model in the study, we believe that other LLMs trained on similar resources, such as the PaLM and the open-sourced Llama model, could also deliver comparable or even better performance*.” Göldi et al. [46] mentioned numerous drawbacks of current GPT-3.5 models but suggested that “*future improvement in these models could mitigate such limitations*.” This acknowledgement could potentially defer the responsibility of ensuring research validity in highly context-dependent HCI studies to LLM model developers.

**Research validity surrounding prompts** is another emerging limitation. *Of the 153 papers, 146 conducted some form of systems or studies that prompted LLMs. Of the 146, 40.4% (N=59) did not release their prompts in the full paper or the supplementary materials.* Authors were generally aware that prompt variation could impact their results: Cheng et al. [22] noted that “*minor prompt adjustments aimed at improving one aspect often had unintended, drastic negative effects on others*.” Similarly, LLM-powered tools, which have been evaluated through technical or user studies, may not generalize externally to other prompts. For example, Kabir et al. [75] mentioned that the design of the prompt in their study is highly dependent on the questions in their sample. Since how people phrase these questions in the real world varies from person to person and situation to situation, more work is needed to evaluate LLMs against prompt variation. Despite the validity concerns, several authors still proposed to revise the prompt to enhance the system. For example, Wang et al. [159] proposed to improve the system quality by adapting their system prompt depending on the input, but acknowledged that this proposal might “*lead to inferior performances*.”

**4.4.4 Consequences** (22.88%, N=35). This category shows potential negative outcomes that may arise from the artifact or study. In some cases, authors present the concerns in an ethics or impact statement (14.38%, N=22) with concrete remediation strategies.

**Economic Harms** (11.11%, N=17) This refers to potential effects on employment and work. For example, De La Torre et al. [32] highlighted the concern of “*developers and creators being replaced*”. However, they also recognized that these tools have not achieved end-to-end development, and if so, these tools should still require human intervention. Shaikh et al. [141] mentioned that their tool to simulate conflict resolution scenarios might cause job replacement and devaluation for expert trainers. Many papers on Communication & Writing, such as Lee et al. [90] and Hoque et al. [62], stated that their LLM-powered writing tool may change copyright issues and how writers work.

**Representational Harms** (5.88%, N=9) This harm refers to social groups being cast in a less favorable light than others, affecting the understandings, beliefs, and attitudes that people hold about these groups [6]. For example, Benharra et al. [11] recognized that LLM-generated personas “*have the potential to reproduce harmful stereotypes*.” Salminen et al. [137] called out that “*as with any novel technology*,” their use of LLM can have adverse societal effects including “*reinforcing gender stereotypes and affecting diversity representation*.” However, these risks were “*not in the scope of*” their study, but warrant further scrutiny from the HCI research community.

**Misinformation Harms** (2.61%, N=4) This harm arises from the LLMs outputting false, misleading, non-sensical, or poor quality information [165]. For example, Li et al. [96] added that writers’ viewpoints may get misled by “*misinformation generated by AI assistants*.” Tanprasert et al. [155] recognized that if they shifted their topic in the study to a more technical topic, which may lead to more cases of LLM hallucination, not only would the research validity have been compromised (i.e., “*the credibility of the information can seriously weaken the chatbot’s stance in the study*”), but users also may suffer from misinformation spread by the chatbot, if the users are not aware of it. We also found studies that tackle misinformation directly in Zhou et al. [186] where they examined characteristics of LLM-generated misinformation compared with human creations, and then evaluated the applicability of existing solutions.

**Malicious Use** (1.96%, N=3) This harm stems from humans intentionally using the LLM to cause harm, e.g., via malware or fraud [165]. Precel et al. [130] used a whole appendix section to discuss how their findings may harm the Jewish community by anti-Semitic actors. When studying the effect of LLM-powered search systems on information-seeking tasks, Sharma et al. [143] recognized that their system and study “*may incur misuse*” because they introduced opinion bias to power the LLM-based search system. Therefore, they “*made public the prompts in the study but will only make them available for requests that we can verify for safe usage (e.g., scientific and non-commercial purposes)*.” This approach highlights the interesting balance between ensuring open source and preventing malicious use.

**Hate Speech** (1.96%, N=3) This category represents prejudice, hostility, or violence against individuals or groups. De La Torre et al. [32] stressed that a serious concern is “*the potential for individuals to generate harmful and inappropriate content*” with their framework, calling for future safeguards. Kim et al. [80] extensively discussed the ethical concerns of using LLMs for personal journaling. They mentioned that their study may suffer from LLMs’ potential “*to generate offensive or violent content*.” To mitigate this risk, the authors informed participants about potential misbehaviors and offered university mental health care resources in case of adverse events.

**Environmental harms** (0.65%, N=1) This category refers to the damage that LLMs can cause the environment, in particular due to the large energy consumption that training and querying requires [165]. One paper explicitly discussed environmental harms [88], in the context of worries to scale up their system with LLMs.

## 5 Discussion

We show substantial growth at CHI in research studying LLMs, echoing trends in other fields [118, 170]. In this section, we discuss where the CHI community has focused its explorations to-date, and what the surge in LLM interest means for HCI's norms around prototyping and design (5.1). We then assess issues regarding research rigor that permeate the field (5.2). We close with a proposal (5.3): a set of *guiding questions* for HCI researchers to reflect on throughout an LLM-powered project, by considering the *task-appropriateness* of their proposed LLM use, the *validity and reproducibility* of their conclusions, and the *consequences* of their work for research participants, technology users, and society.

### 5.1 Revealed Growth Opportunities for HCI

To our best knowledge, our work is the first to systematically characterize how LLMs have influenced HCI research. We find substantial opportunity to expand the application domains where LLMs are used (5.1.1), build theories and methods with lasting impact from the large body of empirical work (5.1.2), and standardize how LLMs can and should impact prototyping practices (5.1.3).

**5.1.1 Beyond language-based applications.** Our study demonstrates that LLMs are applied across a wide range of **application domains**, reflecting diversity within the HCI community, and also confirming that LLMs' influence on HCI is pervasive across subareas. Some areas are well-represented already, and provide examples of how to build new research communities around LLM applications. Specifically, we found the **Communication & Writing** domain has garnered the most attention, perhaps due to LLMs' direct relevance to producing language. This community has coalesced around such initiatives as the In2Writing workshops at NLP and HCI conferences [19], and Lee et al. [90]'s effort to chart the design space of intelligent and interactive writing assistants. Other areas are less represented in our review, and represent opportunities for new research and community-building. For example, papers related to Games and Play were less common in our sample, even though this area is large enough to warrant its own CHI subcommittee. As LLMs facilitate more games and simulations, we anticipate this area to be a generative site for new work. Our categorization of application domains can help researchers identify where in the community their interests might fit, and how to develop these areas as LLMs continue to proliferate.

**5.1.2 Beyond empirical and artifact contributions.** While we observed that HCI researchers succeeded in applying LLMs to a diversity of application domains, we found less diversity in the **contribution types** pursued in the literature. The LLM-related papers in our sample predominantly center on artifact contributions and empirical evaluations, often in the form of user studies of new artifacts. Empiricism is central to understanding phenomena; however, to develop knowledge from our aggregate body of observations, we encourage more attention to Wobbrock and Kientz [168]'s five other contribution types, each of which was less well-represented in the literature.

We observed an opportunity for the community to further pursue **dataset contributions** [168]—and approaches to data collection that

center real user needs and downstream harms. Traditional NLP benchmarks are often criticized for their lack of *context realism*: the model performance measures are often divorced from downstream use cases [100]. Adopting community-driven and participatory approaches to benchmarking could provide data that represents real and diverse user requirements, while still enabling developers to test LLMs' capabilities [151]. HCI's sociotechnical approach is well-positioned to innovate on benchmarking culture: e.g., Bragg et al. have explored how to build automatic sign language translation tools by crowdsourcing video datasets with the Deaf community [13, 156], Reinecke et al. have developed volunteer-based online platform to reach larger and more diverse user population [133, 134], and Kuo et al. [86]'s Wikibench offers a community-driven alternative to data curation that captures community consensus.

We also see significant need for theoretical and methodological contributions, as well as literature review and opinion pieces, all of which can help shape public perception and understanding of LLMs' pitfalls and potential. Theoretical and methodological contributions can draw transferable principles from bodies of empirical and artifact research [10, 168]. Based on our review, the field would benefit from more work on, e.g., the design space around LLMs in the various application domains (cf. Lee et al. [90]'s work on intelligent writing assistants), or the design processes behind developing LLM-based systems. Literature review like the present study and opinion contributions can also help us reflect on our community's progress and help the field to identify and address emergent issues (cf. Correll [28]'s work advocating for moral obligation among researchers and designers in visualization). Our literature review can help fill this void, but more work is needed. Targeted literature review on more specific topics can further guide our community forward, and we especially encourage papers that synthesize lessons bridging HCI and fields like NLP and ML.

**5.1.3 How LLMs impact prototyping standards.** More broadly, our findings signal broader methodological questions for HCI: What level of prototype fidelity is needed to demonstrate a new interaction—and relatedly, what level of system-building and evaluation is needed to make an artifact contribution? This question arises from our challenge to define which papers proposed artifact contributions, and which used **LLMs as system engines**. Throughout HCI, *Wizard of Oz* approaches have long been used to prototype interactions with intelligent agents [30, 115]. These methods typically present a research participant with an interface that appears to have machine intelligence, but unbeknownst to them, a human performs those functions (cf. [50]). Wizard of Oz approaches gained popularity in HCI as methods that allowed rapid and inexpensive prototyping of future technological capabilities.

However, the utility of these methods may change as LLMs proliferate. If a researcher explores the design space around using LLMs in a given domain—to "*sketch with AI*", as Yang et al. [176] describe—does a Wizard of Oz approach provide benefits over a fully automated approach anymore? Historically, researchers have been trained to prototype quickly and cheaply, and thus they might conclude that a Wizard of Oz approach makes more sense. Today, however, LLMs have likely lowered the barrier of developing systems so much that we may expect designers to use them to achieve



more ecologically valid research. After all, even the best of Wizard of Oz methods cannot perfectly proxy machine intelligence [178]. If a designer creates an LLM-backed prototype, however, what level of performance should they aspire to in their system? Should the prototype be evaluated as an artifact contribution? Is it a system contribution, if the implementation is straightforward, given LLMs' capabilities? For HCI, this debate has ramifications not only for the methodological norms we use, teach, and expect in peer review, but also for the research validity that we produce and the contribution that we value. We encourage the HCI community to collectively reflect on what widespread LLM usage means for prototyping standards, and the resulting implications for how HCI produces knowledge.

## 5.2 Challenges: Validity, Reproducibility, and Consequences

To achieve the opportunities we outline in the previous section, the HCI community will also need to reflect on some fundamental challenges with LLM research identified in our analysis.

**5.2.1 Proprietary LLMs raise reproducibility concerns.** Our analysis showed that despite authors' commonly articulated limitations surrounding **research validity**, papers using LLMs is growing exponentially. This trend adds urgency to calls for examining reproducibility in HCI [26, 136].

We found that an overwhelming majority of our sample (84.98%) studied a variation of the closed GPT-family models ( $N_{\text{GPT-4}} = 61$ ,  $N_{\text{GPT-3.5}} = 41$ ,  $N_{\text{GPT-3}} = 26$ ). Using closed models can pose serious problems for research reproducibility, especially if authors choose not to disclose prompts. Researchers have shown that proprietary and closed models, often accessed through APIs, are constantly changing, and may even inject unspecified edits to a user's prompt (e.g., system prompts) [135], meaning model behavior may be unpredictable despite using the same, disclosed prompt. Proprietary models also generally do not disclose their training data or model weights, which makes understanding model behaviors and properties in applications and downstream systems difficult. Most papers also did not justify their model choice, though model families can exhibit quite different behaviors. For example, models optimized for chat and those optimized to take instructions can be expected to behave differently [122, 183]. In our study, we found few authors explicitly justified their model choice for their use case. Some artifact contribution papers implied that their systems could be *modular*, suggesting the LLMs in the system could be customized by users or in future work. If an LLM-powered system is meant to be modular, and models exhibit different behaviors, then developers should disclose and discuss how model choices might affect the system for future users and developers. To further complicate the issue, 40.41% of the papers ( $N = 59$ ) did not release the prompts in the paper or supplementary material. Given LLMs are sensitive to subtle changes in prompt formatting even in large models [139], the lack of transparency in prompt design and usage may affect system performance, and prevent researchers from replicating and building upon existing work.

**5.2.2 LLM properties introduce additional research validity concerns.** Our analysis surfaced the fact authors have many concerns

around how LLMs' inherent properties might impact research validity—but less knowledge about what precisely to do about it. Whether LLMs were the object of study or powered a system with which users interact, researchers readily acknowledged issues like **LLMs' inherently limited training data**, penchant for **hallucination**, and **nondeterministic responses**. Some of these limitations shaped whether we *should* expect certain behaviors from LLMs (e.g., limitations on training data), and others shaped whether research results could be considered externally valid (e.g., hallucination and nondeterminism).

However, we found that the most commonly mentioned LLM-related performance limitations were **unspecified errors and biases** (16%,  $N=24$ ). Though authors have some awareness of LLMs' limitations, this code's prevalence indicates that further engagement with the precise nature and performance effects of these errors was often unaddressed. Being more specific about the nature of errors or bias arising from LLMs and how this may affect the system or results is critical for a reader's understanding of the nature and extent of the stated limitation. For example, the more specific issues captured by other codes, e.g., "hallucination," bring with them the ability to better interpret specific potential failure types of a system, and even imagine potential downstream harms. We urge HCI researchers to more precisely specify what potential errors and biases they identify in their use of LLMs, so that consumers of our research can better understand how the systems built upon these technologies may fail.

**5.2.3 Consequences, Risks, and Broader Impacts.** In parallel to the *limitations* around validity and reproducibility described in the previous section, we found tremendous need to confront how HCI researchers assess and report the *consequences and risks* of their work.

First, we explicitly differentiate between *limitations* and *consequences*. *Limitations* refer to factors that affect the truthfulness of the paper's conclusions, such as issues with validity, transferability, and generalizability. *Consequences* pertain to long-term social impact, including insights that could help guide real-world deployments. In fields such as ML/AI and computer security, recent initiatives have asked authors to provide *ethics statements* [59], *broader impact statements* [119], and other structured ways of reflecting on the consequences of their work.

While authors considered questions of validity and reproducibility—limitations of their work—only 35 papers discussed potential consequences of their findings and results, often in an ethics statement ( $N=22$ ). Ethics statements were discussed among HCI researchers in 2018 [59], but to-date have not been formally standardized in CHI's submission process; however, they have been used in ML and AI conferences including NeurIPS and FAccT [119]. As our study showed that LLMs have been used in diverse applications and are changing research practices, we believe that the CHI community should place greater emphasis on discussions around consequences. Encouraging a more explicit discussion ensures that the HCI contributions are responsibly aligned with the broader societal good.

More broadly, we contend that structured consideration of consequences — via an ethics statement or other means — would help HCI lead the scientific community by demonstration as LLM-based



work proliferates. As HCI research inherently considers people and society, its innovations are likely to be deployed and have impact with real-world users [27]. Establishing field norms to consider consequences can help HCI lead in engaging with LLMs in rigorous and thoughtful ways, providing a model for researchers and practitioners across disciplines. Below, we contribute an initial proposal we hope accelerates the scientific community towards this vision.

### 5.3 Guiding questions for HCI researchers using LLMs

Given serious concerns around validity, reproducibility, risks and consequences (5.2), how can we move towards the opportunities outlined in 5.1? As a first step, we contribute practical guidance to prompt researchers' reflection on the *validity* and *appropriateness* of their LLM-related work. We view LLM-related research not as categorically harmful, but rather that using LLMs requires careful consideration throughout a project.

In this section, we draw on the **LLM roles** and **Limitations** to synthesize *guiding questions* for researchers and practitioners to consider at each stage of the research process, and for each role that an LLM might take. Our questions are intended as prompts for reflection, not a checklist for completion, and should be used iteratively to ensure that any work continually centers thoughtful LLM usage. We chose more open-ended questions over prescriptive guidance to uplift critical thinking around LLM usage in research design. **G** represents *general* questions for any project, whereas **S** refers to *specific* questions for each role.

1. **G** **What role will the LLM play in your project?** A researcher should understand the stage at which an LLM might be included (Figure 3). A key question is whether an LLM is needed at all—whether achieving the same result is possible using alternate approaches that are better established or less costly, such as using simpler models or humans.

2. **G** **Which model is appropriate?** This question helps authors decide between open and closed models. As Palmer et al. [125] discuss, closed models are at odds with the transparency and reproducibility expected of research. Using closed and proprietary LLMs may also violate study ethics if participants have not consented to sharing data with the LLM. Still, closed models may be appropriate if, e.g., LLMs are the object of study (as in an audit study); if cheap and rapid prototypes are needed; or if a closed model was shown to be the state of the art in a specific task, and is used only for that task. If others may treat the LLM in a system as modular, then consider the robustness of the chosen model and the impact of swapping in different models. Researchers should consider such factors and justify their model choice.

3. **G** **How did you disclose the models and prompts?** This question encourages authors to document model versions and prompts used in their study. For models, we encourage specificity: e.g., gpt-4o-2024-08-06 and gpt-4o can refer to different models. Authors should also clearly document the full prompts, or the prompt templates provided to users. Authors can also consider other methods to improve the research validity, e.g., fine-tuning an open source model on domain-specific datasets [64, 164].

4. **G** **What are the potential limitations of using LLMs for your selected role?** For each LLM role, we contribute specific sets of reflective questions.

#### • LLMs as system engines:

- S** *What level of artifact fidelity is appropriate to support the contribution?* If the main contribution is a formative study or user perceptions of specific LLM outputs, enabling the interaction is perhaps more important than deploying a fully-functional system. A Wizard of Oz approach may be more appropriate than building a system around a commercial LLM API.
- S** *How would factors like models and prompts affect the system performance?* This clarifies whether the system can achieve the claimed effectiveness with different models or changes in prompts.
- S** *How would factors like models and prompts used in the system affect the research validity of the user study?* Authors should consider whether the LLM-powered system is robust across users, and note any discrepancies between target system users and recruitment population.

#### • LLMs as research tools:

- S** *Why are LLMs appropriate for your research task?* If your task is *classification*, e.g., labeling a dataset, using an LLM may overlook nuances in the human *perspectives*. If your task is *generation*, e.g., creating survey questions or datasets, using an LLM risks neglecting lay and domain *expertise*.
- S** *How can you evaluate the performance of your LLM-based research tool?* Across tasks, validation via human or formal methods is often needed to quality-check an LLM's outputs. These evaluations are vital, but the human effort needed to structure and faithfully execute them may exceed the utility of using the LLM in the first place—what Bainbridge [5] calls the “automation trap”.
- S** *How will the performance of your LLM-powered research tool affect the validity of your research?* In addition to ensuring research tools remain standardized and accurate, authors should understand how the choice to use an LLM would affect the claims of the empirical work.

#### • LLMs as participants & users. Consider the questions under **LLMs as research tools** above. Then, specifically:

- S** *Given LLMs' known inability to faithfully represent people, how can an LLM-powered tool adequately stand in for the target population in your study?* Using LLMs to simulate users deprives them of the opportunity to consent to such research [1]. LLMs also run the risk of misrepresenting people and are unlikely to faithfully portray identity groups due to the nature of their training data [158]. Given these known constraints, consider how to adjust your study design to enable people from your target population to evaluate the LLM's outputs, and determine how they are used (cf. [151, 153]). Throughout, stay attuned to whether the effort required for proper human evaluation and participation exceeds the benefits of introducing an LLM in the first place.

- Given that they are only trained on human language, how can LLM-backed tools reflect the realism of human behavior and opinion dynamics of interest? Although LLMs might display human-like behaviors and opinion dynamics by modeling language, they often struggle to generate outputs that capture the complexity and diversity of real human interactions shaped by individuals' lived experiences. Human language is also inherently limited in capturing the fidelity of human behavior, which can threaten this method's validity. Given these known constraints, consider how to adjust your study design to evaluate the LLM's outputs against real discourse and deliberation (cf. [86]). Again, consider whether the effort required for proper human evaluation might exceed the benefits of introducing an LLM.

- **LLMs as objects of study:**

- What model behaviors do you aim to study? When studying LLMs directly, authors should consider a common feature of most LLMs, a feature of particular model families, or just one particular model (e.g., gpt-4o-2024-08-06).
- How did you ensure that the claims made about the models were appropriate? Authors should consider whether the findings can generalize to other models. If not, authors should not overclaim the findings.

- **User perceptions of LLMs:**

- Who are the representative participants for the study? Authors should consider how their participants impact the internal validity of their work: whether their study sample accurately reflects the population they claim to represent.
- What confounds could impact participants' perceptions of LLMs? LLMs are subject to tremendous hype in the popular press. Participants may come with preconceptions about LLMs' capabilities that require researchers' attention. For example, a participant who has seen ads from AI companies may more quickly grasp the affordances of a new LLM-powered interaction paradigm than a participant who does not experience AI filter bubbles.

##### 5. **What are the potential consequences of your study?**

LLMs have known *environmental* costs authors should consider in the study design (cf. [110]). Having participants interact with LLMs may also impact *privacy* [17], especially when using closed models; thus authors may consider how to obtain consent for an LLM to use a participant's data, how to sanitize LLM inputs, and measures to protect participants' agency over their data. HCI researchers studying LLMs—especially when they augment or replace human effort—should consider the systems' *economic* impacts. LLMs' need for massive datasets can create global inequalities for data workers [48]; and companies may prioritize investing in LLMs over workers, even as humans are needed to ensure LLMs function properly [5].

## 5.4 Limitations

Our work has several limitations. First, our sampling approach might not cover all papers that used LLMs. For instance, we found one paper in our robustness check that mentioned GPT-4 just once, in their methods, without mentioning any other keywords in our list. Other works may have even used LLMs in their methods without

mentioning them at all, which would align with the increasing interest in using LLMs to automate academic research [109]. Our work primarily focused on prompting as the main interface, but future study may extend our samples to study and identify best practice for other techniques (e.g., fine-tuning [55], LLM-based embeddings [129], and multi-agents [51]). While insights from this paper (e.g., computational cost) remain relevant, additional research validity concerns may emerge, e.g., challenges in sharing datasets to replicate fine-tuning results or agent configurations to reproduce multi-agent system outcomes. We may also miss CHI papers that did not pass our keyword filters, e.g., [77]. To remedy these limitations, we release our paper annotation as a living artifact to encourage community to extend our dataset and annotations.

Second, our review was limited in scope by the manual and iterative process. Using LLMs to conduct analyses like ours is an active research topic and can increase scale, but we chose not to use LLMs because of concerns with using LLMs without proper disclosure and evaluation. In our preliminary phase, we used gpt-4o-2024-05-13 to explore paper topics, but found the themes too general to gain meaningful insights. To use LLMs effectively, human qualitative coding will likely still be required to develop effective prompts and validate the accuracy of the LLM classifier. Hence, we spent hours curating the dataset, reading papers, resolving coding disagreements, and discussing difficult papers. The laborious nature of this process prevented us from conducting a more expansive literature review. Future researchers might use our paper as a starting point to examine LLMs' impact on other HCI subcommunities and even conferences in other fields, e.g., by reviewing the (dis-)connection between the NLP and HCI communities.

## Acknowledgments

We thank the anonymous reviewers for their valuable feedback, and the ACM Digital Library for supplying the data for our study. We also thank Katharina Reinecke, Jenn Wortman Vaughan, Mert Yuksekgonul, Kevin Feng, Mohammed Alsobay, Sachita Nishal, Harsh Kumar, Shivani Kapania, and many more friends and mentors at the University of Washington and Microsoft Research for their research inspirations, fun conversations, and helpful suggestions.

## References

- [1] William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Diaz, Selim El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. 2024. The Illusion of Artificial Inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 286, 12 pages. <https://doi.org/10.1145/3613904.3642703>
- [2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300233>
- [3] Ian Arawajo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 304, 18 pages. <https://doi.org/10.1145/3613904.3642016>
- [4] Marianne Aubin Le Quéré, Hope Schroeder, Casey Randazzo, Jie Gao, Ziv Epstein, Simon Tangi Perrault, David Mimno, Louise Barkhuus, and Hanlin Li. 2024. LLMs as Research Tools: Applications and Evaluations in HCI Data Work. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA). Association for Computing Machinery, Article 10, 10 pages.

- Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 479, 7 pages. <https://doi.org/10.1145/3613905.3636301>
- [5] Lisanne Bainbridge. 1983. Ironies of automation. *Automatica* 19, 6 (1983), 775–779. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)
  - [6] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society*. New York, NY, 1.
  - [7] Nikki M Barrington, Nithin Gupta, Basel Musmar, David Doyle, Nicholas Panico, Nikhil Godbole, Taylor Reardon, and Randy S D'Amico. 2023. A bibliometric analysis of the rise of ChatGPT in medical research. *Medical Sciences* 11, 3 (2023), 61.
  - [8] Christoph Bartneck and Jun Hu. 2009. Scientometric analysis of the CHI proceedings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) (CHI '09). Association for Computing Machinery, New York, NY, USA, 699–708. <https://doi.org/10.1145/1518701.1518810>
  - [9] Yasmine Belghith, Atefeh Mahdavi Goloujeh, Brian Magerko, Duri Long, Tom McKlin, and Jessica Roberts. 2024. Testing, Socializing, Exploring: Characterizing Middle Schoolers' Approaches to and Conceptions of ChatGPT. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 276, 17 pages. <https://doi.org/10.1145/3613904.3642332>
  - [10] Victoria Bellotti, Maribeth Back, W. Keith Edwards, Rebecca E. Grinter, Austin Henderson, and Cristina Lopes. 2002. Making sense of sensing systems: five questions for designers and researchers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA) (CHI '02). Association for Computing Machinery, New York, NY, USA, 415–422. <https://doi.org/10.1145/503376.503450>
  - [11] Karim Benharrah, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-Defined AI Personas for On-Demand Feedback Generation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1049, 18 pages. <https://doi.org/10.1145/3613904.3642406>
  - [12] Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics* 5, 5 (2023), 277–280.
  - [13] Danielle Bragg, Abraham Glasser, Fyodor Minakov, Naomi Caselli, and William Thies. 2022. Exploring collection of sign language videos through crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–24.
  - [14] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 732, 13 pages. <https://doi.org/10.1145/3411764.3445372>
  - [15] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (CHI '16). Association for Computing Machinery, New York, NY, USA, 981–992. <https://doi.org/10.1145/2858036.2858498>
  - [16] Hancheng Cao, Yujie Lu, Yuting Deng, Daniel McFarland, and Michael S. Bernstein. 2023. Breaking Out of the Ivory Tower: A Large-scale Analysis of Patent Citations to HCI Research. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 760, 24 pages. <https://doi.org/10.1145/3544548.3581108>
  - [17] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying Memorization Across Neural Language Models. arXiv:2202.07646 [cs.LG] <https://arxiv.org/abs/2202.07646>
  - [18] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or Artifice? Large Language Models and the False Promise of Creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 30, 34 pages. <https://doi.org/10.1145/3613904.3642731>
  - [19] Minsuk Chang, John Joon Young Chung, Katy Ilonka Gero, Ting-Hao Kenneth Huang, Dongyeop Kang, Vipul Raheja, Sarah Sterman, and Thiemo Wambgsans. 2024. Dark Sides: Envisioning, Understanding, and Preventing Harmful Effects of Writing Assistants - The Third Workshop on Intelligent and Interactive Writing Assistants. , 6 pages. <https://doi.org/10.1145/3613905.3636312>
  - [20] Chaoran Chen, Weijun Li, Wenxin Song, Yanfang Ye, Yaxing Yao, and Toby Jia-Jun Li. 2024. An Empathy-Based Sandbox Approach to Bridge the Privacy Gap among Attitudes, Goals, Knowledge, and Behaviors. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 234, 28 pages. <https://doi.org/10.1145/3613904.3642363>
  - [21] John Chen, Xi Lu, Yuzhou Du, Michael Rejtig, Ruth Bagley, Mike Horn, and Uri Wilensky. 2024. Learning Agent-based Modeling with LLM Companions: Experiences of Novices and Experts Using ChatGPT & NetLogo Chat. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 141, 18 pages. <https://doi.org/10.1145/3613904.3642377>
  - [22] Alan Y. Cheng, Meng Guo, Melissa Ran, Arpit Ranasarria, Arjun Sharma, Anthony Xie, Khuyen N. Le, Bala Vinaitirthan, Shihe (Tracy) Luan, David Thomas Henry Wright, Andrea Cuadra, Roy Pea, and James A. Landay. 2024. Scientific and Fantastical: Creating Immersive, Culturally Relevant Learning Experiences with Augmented Reality and Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 275, 23 pages. <https://doi.org/10.1145/3613904.3642041>
  - [23] Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. 2024. RELIC: Investigating Large Language Model Responses using Self-Consistency. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 647, 18 pages. <https://doi.org/10.1145/3613904.3641904>
  - [24] Madiha Zahrah Choksi, Marianne Aubin Le Quéré, Travis Lloyd, Ruojia Tao, James Grimmelmann, and Mor Naaman. 2024. Under the (neighbor)hood: Hyperlocal Surveillance on Nextdoor. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 771, 22 pages. <https://doi.org/10.1145/3613904.3641967>
  - [25] John Joon Young Chung, Woosok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 209, 19 pages. <https://doi.org/10.1145/3491102.3501819>
  - [26] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. HARK No More: On the Preregistration of CHI Experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3173715>
  - [27] Lucas Colusso, Ridley Jones, Sean A. Munson, and Gary Hsieh. 2019. A Translational Science Model for HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300231>
  - [28] Michael Correll. 2019. Ethical Dimensions of Visualization Research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300418>
  - [29] Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F. Jung, Nicola Dell, Deborah Estrin, and James A. Landay. 2024. The Illusion of Empathy? Notes on Displays of Emotion in Human-Computer Interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 446, 18 pages. <https://doi.org/10.1145/3613904.3642336>
  - [30] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st International Conference on Intelligent User Interfaces* (Orlando, Florida, USA) (IUI '93). Association for Computing Machinery, New York, NY, USA, 193–200. <https://doi.org/10.1145/169891.169968>
  - [31] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 408, 17 pages. <https://doi.org/10.1145/3544548.3580969>
  - [32] Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. 2024. LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 600, 22 pages. <https://doi.org/10.1145/3613904.3642579>
  - [33] Nicola Dell and Neha Kumar. 2016. The Ins and Outs of HCI for Development. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (CHI '16). Association for Computing Machinery, New York, NY, USA, 2220–2232. <https://doi.org/10.1145/2858036.2858081>
  - [34] Xiaohan Ding, Buse Carik, Uma Sushmitha Gunturi, Valerie Reyna, and Eugenia Ha Rim Rho. 2024. Leveraging Prompt-Based Large Language Models: Predicting Pandemic Health Decisions and Outcomes Through Social Media Language. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 443, 20 pages. <https://doi.org/10.1145/3613904.3642117>
  - [35] Kimberly Do\*, Rock Yuren Pang\*, Jiachen Jiang, and Katharina Reinecke. 2023. "That's important, but...": How Computer Science Researchers Anticipate Unintended Consequences of Their Research Innovations. In *Proceedings of the 2023*

- CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 602, 16 pages. <https://doi.org/10.1145/3544548.3581347>
- [36] Peitong Duan, Jeremy Warner, Yang Li, and Bjoern Hartmann. 2024. Generating Automatic Feedback on UI Mockups with Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 6, 20 pages. <https://doi.org/10.1145/3613904.3642782>
- [37] Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. 2024. A Bibliometric Review of Large Language Models Research from 2017 to 2023. *ACM Trans. Intell. Syst. Technol.* (may 2024). <https://doi.org/10.1145/3664930> Just Accepted.
- [38] Li Feng, Ryan Yen, Yuzhe You, Mingming Fan, Jian Zhao, and Zhicong Lu. 2024. Coprompt: Supporting prompt sharing and referring in collaborative natural language programming. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.
- [39] Sidong Feng, Suyu Ma, Han Wang, David Kong, and Chunyang Chen. 2024. MUD: Towards a Large-Scale and Noise-Filtered UI Dataset for Modern Style UI Modeling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.
- [40] Raymond Fok, Nedim Lipka, Tong Sun, and Alexa F Siu. 2024. Marco: Supporting Business Document Workflows via Collection-Centric Information Foraging with Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.
- [41] Liye Fu, Benjamin Newman, Maurice Jakesch, and Sarah Kreps. 2023. Comparing sentence-level suggestions to message-level suggestions in AI-mediated communication. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [42] Yue Fu, Sami Foell, Xuhai Xu, and Alexis Hiniker. 2024. From Text to Self: Users' Perception of AIMC Tools on Interpersonal Communication and Self. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 977, 17 pages. <https://doi.org/10.1145/3613904.3641955>
- [43] Takao Fujii, Katie Seaborn, and Madeleine Steeds. 2024. Silver-tongued and sundry: Exploring intersectional pronouns with chatgpt. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.
- [44] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).
- [45] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. 2023. Social Dynamics of AI Support in Creative Writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 245, 15 pages. <https://doi.org/10.1145/3544548.3580782>
- [46] Andreas Göldi, Thimo Wambsganss, Seyed Parsa Neshaei, and Roman Rietsche. 2024. Intelligent Support Engages Writers Through Relevant Cognitive Processes. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–12.
- [47] Alice Good and Arunasalam Sambhanthan. 2014. Accessing web based health care and resources for mental health: interface design considerations for people experiencing mental illness. In *Design, User Experience, and Usability: User Experience Design for Everyday Life Applications and Services: Third International Conference, DUXU 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part III* 3. Springer, 25–33.
- [48] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.
- [49] Jonathan Grudin. 2009. AI and HCI: Two fields divided by a common focus. *AI magazine* 30, 4 (2009), 48–48.
- [50] Ken Gu, Ruoxi Shang, Tim Althoff, Chenglong Wang, and Steven M Drucker. 2024. How Do Analysts Understand and Verify AI-Assisted Data Analyses?. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.
- [51] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* (2024).
- [52] Justin Anthony Haegele and Samuel Hodge. 2016. Disability discourse: Overview and critiques of the medical and social models. *Quest* 68, 2 (2016), 193–206.
- [53] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [54] Ariel Han, Xiaofei Zhou, Zhenyao Cai, Shenshen Han, Richard Ko, Seth Corrigan, and Kylie A Peppler. 2024. Teachers, Parents, and Students' perspectives on Integrating Generative AI into Elementary Literacy Education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [55] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608* (2024).
- [56] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (01 2020), 89–100. <https://doi.org/10.1093/jcmc/zmz022> arXiv:<https://academic.oup.com/jcmc/article-pdf/25/1/89/32961176/zmz022.pdf>
- [57] Yuexing Hao, Zeyu Liu, Robert N. Riter, and Saleh Kalantari. 2024. Advancing Patient-Centered Shared Decision-Making with AI Systems for Older Adult Cancer Patients. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 437, 20 pages. <https://doi.org/10.1145/3613904.3642353>
- [58] Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1040, 25 pages. <https://doi.org/10.1145/3613904.3642834>
- [59] Brent Hecht, Lauren Wilcox, Jeffrey P. Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Luigi De Russis, Lana Yarosh, Bushra Anjum, Danish Contractor, and Cathy Wu. 2021. It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process. arXiv:2112.09544 [cs.CY] <https://arxiv.org/abs/2112.09544>
- [60] Michael A. Hedderich, Natalie N. Bazarova, Wenting Zou, Ryun Shim, Xinda Ma, and Qian Yang. 2024. A Piece of Theatre: Investigating How Teachers Design LLM Chatbots to Assist Adolescent Cyberbullying Education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 668, 17 pages. <https://doi.org/10.1145/3613904.3642379>
- [61] Hendrik Heuer and Daniel Buschek. 2021. Methods for the Design and Evaluation of HCI+NLP Systems. In *Proceedings of the First Workshop on Bridging Human-Computer Interaction and Natural Language Processing*, Su Lin Blodgett, Michael Madaio, Brendan O'Connor, Hanna Wallach, and Qian Yang (Eds.). Association for Computational Linguistics, Online, 28–33. <https://aclanthology.org/2021.hcinlp-1.5>
- [62] Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia D. Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist. 2024. The HaLLMark Effect: Supporting Provenance and Transparent Use of Large Language Models in Writing with Interactive Visualization. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1045, 15 pages. <https://doi.org/10.1145/3613904.3641895>
- [63] Yihan Hou, Manling Yang, Hao Cui, Lei Wang, Jie Xu, and Wei Zeng. 2024. C2Ideas: Supporting Creative Interior Color Design Ideation with a Large Language Model. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 172, 18 pages. <https://doi.org/10.1145/3613904.3642224>
- [64] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
- [65] Forrest Huang, Gang Li, Tao Li, and Yang Li. 2024. Automatic Macro Mining from Interaction Traces at Scale. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.
- [66] Rong Huang, Haichuan Lin, Chuanshang Chen, Kang Zhang, and Wei Zeng. 2024. PlantoGraphy: Incorporating iterative design process into generative artificial intelligence for landscape rendering. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.
- [67] Jessica Hullman. 2024. Status update on Twitter. <https://x.com/JessicaHullman/status/1791645453223608422> Accessed: 2024-07-15.
- [68] Gautier Izacard and Édouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 874–880.
- [69] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–15.
- [70] JiWoong Jiang, Sanika Moharana, Patrick Carrington, and Andrew Begel. 2024. "It's the only thing I can trust": Envisioning Large Language Model Use by Autistic Workers for Communication Assistance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.
- [71] Ellen Jiang, Edwin Toh, Alejandra Molina, Kristen Olson, Claire Kayacik, Aaron Donsbach, Carrie J Cai, and Michael Terry. 2022. Discovering the Syntax and Strategies of Natural Language Programming with Generative Language Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 386, 19 pages. <https://doi.org/10.1145/3491102.3501870>



- [72] Hyounghook Jin, Seonghee Lee, Hyungyu Shin, and Juho Kim. 2024. Teach AI How to Code: Using Large Language Models as Teachable Agents for Programming Education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–28.
- [73] Eunkyoung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 18, 16 pages. <https://doi.org/10.1145/3544548.3581503>
- [74] Mirabelle Jones, Christina Neumayer, and Irina Shklovski. 2023. Embodying the Algorithm: Exploring Relationships with Large Language Models Through Artistic Performance. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 654, 24 pages. <https://doi.org/10.1145/3544548.3580885>
- [75] Samia Kabir, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2024. Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 935, 17 pages. <https://doi.org/10.1145/3613904.3642596>
- [76] Shivani Kapania, Ruiyi Wang, Toby Jia-Jun Li, Tianshi Li, and Hong Shen. 2024. "I'm categorizing LLM as a productivity tool": Examining ethics of LLM use in HCI research practices. *arXiv:2403.19876 [cs.HC]* <https://arxiv.org/abs/2403.19876>
- [77] Majeed Kazemitabaar, Justin Chow, Carl Ka To Ma, Barbara J. Ericson, David Weintrop, and Tovi Grossman. 2023. Studying the effect of AI Code Generators on Supporting Novice Learners in Introductory Programming. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 455, 23 pages. <https://doi.org/10.1145/3544548.3580919>
- [78] Jin K. Kim, Michael Chua, Mandy Rickard, and Armando Lorenzo. 2023. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology* 19, 5 (2023), 598–604. <https://doi.org/10.1016/j.jpurol.2023.05.018>
- [79] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.
- [80] Taewan Kim, Donghoon Shin, Young-Ho Kim, and Hwajung Hong. 2024. DiaryMate: Understanding User Perceptions and Experience in Human-AI Collaboration for Personal Journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.
- [81] Tae Soo Kim, DaEun Choi, Yoonseo Choi, and Juho Kim. 2022. Stylette: Styling the Web with Natural Language. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 5, 17 pages. <https://doi.org/10.1145/3491102.3501931>
- [82] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 306, 21 pages. <https://doi.org/10.1145/3613904.3642216>
- [83] Hyung-Kwon Ko, Hyeon Jeon, Gwanmo Park, Dae Hyun Kim, Nam Wook Kim, Juho Kim, and Jinwook Seo. 2024. Natural language dataset generation framework for visualizations powered by large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.
- [84] Charlotte Kobiella, Yarhy Said Flores López, Franz Waltenberger, Fiona Draxler, and Albrecht Schmidt. 2024. "If the Machine Is As Good As Me, Then What Use Am I?"—How the Use of ChatGPT Changes Young Professionals' Perception of Productivity and Accomplishment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.
- [85] Emily Kuang, Minghao Li, Mingming Fan, and Kristen Shinohara. 2024. Enhancing UX Evaluation Through Collaboration with Conversational AI Assistants: Effects of Proactive Dialogue and Timing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 3, 16 pages. <https://doi.org/10.1145/3613904.3642168>
- [86] Tzu-Sheng Kuo, Aaron Lee Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. 2024. Wikibench: Community-driven data curation for ai evaluation on wikipedia. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.
- [87] Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLoM. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 766, 28 pages. <https://doi.org/10.1145/3613904.3642830>
- [88] Lane Lawley and Christopher Macellan. 2024. VAL: Interactive Task Learning with GPT Dialog Parsing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 5, 18 pages. <https://doi.org/10.1145/3613904.3641915>
- [89] Jungeun Lee, Suwon Yoon, Kyoosik Lee, Eunae Jeong, Jae-Eun Cho, Wonjeong Park, Dongsun Yim, and Inseok Hwang. 2024. Open Sesame? Open Salami! Personalizing Vocabulary Assessment-Intervention for Children via Pervasive Profiling and Bespoke Storybook Generation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–32.
- [90] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, Antonette Shibani, Disha Shrivastava, Lila Shroff, Agnia Sergeyuk, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia Ha Rim Rho, Zejiang Shen, and Pao Siangliulue. 2024. A Design Space for Intelligent and Interactive Writing Assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1054, 35 pages. <https://doi.org/10.1145/3613904.3642697>
- [91] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.
- [92] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746* (2022).
- [93] Yoonjoo Lee, Hyeonsu B. Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. PaperWeaver: Enriching Topical Paper Alerts by Contextualizing Recommended Papers with User-collected Papers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.
- [94] Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2024. Hill: A hallucination identifier for large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–13.
- [95] Jiahao Nick Li, Yan Xu, Tovi Grossman, Stephanie Santosa, and Michelle Li. 2024. OmniActions: Predicting Digital Actions in Response to Real-World Multimodal Sensory Inputs with LLMs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.
- [96] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. 2024. The Value, Benefits, and Concerns of Generative AI-Powered Assistance in Writing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1048, 25 pages. <https://doi.org/10.1145/3613904.3642625>
- [97] Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2023. Guiding Large Language Models via Directional Stimulus Prompting. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 62630–62656. [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/c5601d99ed028448f29d1dae2e4a926d-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/c5601d99ed028448f29d1dae2e4a926d-Paper-Conference.pdf)
- [98] Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. 2024. Mapping the Increasing Use of LLMs in Scientific Papers. *arXiv:2404.01268 [cs.CL]* <https://arxiv.org/abs/2404.01268>
- [99] Q. Vera Liao, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. 2023. Designing Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 9, 21 pages. <https://doi.org/10.1145/3544548.3580652>
- [100] Q. Vera Liao and Ziang Xiao. 2023. Rethinking Model Evaluation as Narrowing the Socio-Technical Gap. *arXiv:2306.03100 [cs.HC]* <https://arxiv.org/abs/2306.03100>
- [101] David Chuan-En Lin and Nikolas Martelaro. 2024. Jigsaw: Supporting Designers to Prototype Multimodal Applications by Chaining AI Foundation Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 4, 15 pages. <https://doi.org/10.1145/3613904.3641920>
- [102] Susan Lin, Jeremy Warner, J.D. Zamfirescu-Pereira, Matthew G Lee, Sauhard Jain, Shanqing Cai, Piyawat Lertvittayakumjorn, Michael Xuelin Huang, Shumin Zhai, Bjoern Hartmann, and Can Liu. 2024. Rambler: Supporting Writing

- With Speech via LLM-Assisted Gist Manipulation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Vol. 22. ACM, 1–19. <https://doi.org/10.1145/3613904.3642217>
- [103] Yupeng Lin and Zhonggen Yu. 2024. A bibliometric analysis of artificial intelligence chatbots in educational contexts. *Interactive Technology and Smart Education* 21, 2 (2024), 189–213.
- [104] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 143, 14 pages. <https://doi.org/10.1145/3411764.3445488>
- [105] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D Gordon. 2023. “What it wants me to say”: Bridging the abstraction gap between end-user programmers and code-generating large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–31.
- [106] Xingyu “Bruce” Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang “Anthony” Chen, and Ruofei Du. 2023. Visual captions: augmenting verbal communication with on-the-fly visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [107] Xingyu Bruce Liu, Jiahao Nick Li, David Kim, Xiang “Anthony” Chen, and Ruofei Du. 2024. Human I/O: Towards a Unified Approach to Detecting Situational Impairments. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.
- [108] Yong Liu, Jorge Goncalves, Denzil Ferreira, Bei Xiao, Simo Hosio, and Vassilis Kostakos. 2014. CHI 1994–2013: mapping two decades of intellectual progress through co-word analysis. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 3553–3562.
- [109] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv:2408.06292 [cs.AI]* <https://arxiv.org/abs/2408.06292>
- [110] Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2024. Power hungry processing: Watts driving the cost of AI deployment?. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 85–99.
- [111] Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z Gajos. 2024. Evaluating the Experience of LGBTQ+ People Using Large Language Model Based Chatbots for Mental Health Support. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.
- [112] Kelly Mack, Emma McDonnell, Dhruv Jain, Lucy Lu Wang, Jon E. Froehlich, and Leah Findlater. 2021. What do we mean by “accessibility research”? A literature survey of accessibility papers in CHI and ASSETS from 1994 to 2019. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [113] I Scott MacKenzie. 2024. Human-computer interaction: An empirical research perspective. (2024).
- [114] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. Directgpt: A direct manipulation interface to interact with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.
- [115] David Maulsby, Saul Greenberg, and Richard Mander. 1993. Prototyping an intelligent agent through Wizard of Oz. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 277–284.
- [116] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–34.
- [117] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, Prisma Group, et al. 2010. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *International journal of surgery* 8, 5 (2010), 336–341.
- [118] Rajiv Movva, Sidhika Balachandrar, 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.
- [119] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Unpacking the expressed consequences of AI research in broader impact statements. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 795–806.
- [120] Sydney Nguyen, Hannah McLean Babe, Yangtian Zi, Arjun Guha, Carolyn Jane Anderson, and Molly Q Feldman. 2024. How Beginning Programmers and Code LLMs (Mis) read Each Other. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.
- [121] Rajvardhan Oak and Zubair Shafiq. 2024. Understanding Underground Incentivized Review Services. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.
- [122] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.
- [123] Shuyin Ouyang, Jie M Zhang, Mark Harman, and Meng Wang. 2023. LLM is Like a Box of Chocolates: the Non-determinism of ChatGPT in Code Generation. *arXiv preprint arXiv:2308.02828* (2023).
- [124] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj* 372 (2021).
- [125] Alexis Palmer, Noah A Smith, and Arthur Spirling. 2024. Using proprietary language models in academic research requires explicit justification. *Nature Computational Science* 4, 1 (2024), 2–3.
- [126] Rock Yuren Pang, Sebastin Santy, René Just, and Katharina Reinecke. 2024. BLIP: Facilitating the Exploration of Undesirable Consequences of Digital Technologies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 290, 18 pages. <https://doi.org/10.1145/3613904.3642054>
- [127] Hyanghee Park and Daehwan Ahn. 2024. The Promise and Peril of ChatGPT in Higher Education: Opportunities, Challenges, and Design Implications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 271, 21 pages. <https://doi.org/10.1145/3613904.3642785>
- [128] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–16.
- [129] Alina Petukhova, Joao P Matos-Carvalho, and Nuno Fachada. 2024. Text clustering with LLM embeddings. *arXiv preprint arXiv:2403.15112* (2024).
- [130] Heila Precel, Allison McDonald, Brent Hecht, and Nicholas Vincent. 2024. A Canary in the AI Coal Mine: American Jews May Be Disproportionately Harmed by Intellectual Property Disposition in Large Language Model Training. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [131] Mirjana Prpa, Giovanni Maria Troiano, Matthew Wood, and Yvonne Coady. 2024. Challenges and Opportunities of LLM-Based Synthetic Personae and Data in HCI. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 461, 5 pages. <https://doi.org/10.1145/3613905.3636293>
- [132] John Pruitt and Jonathan Grudin. 2003. Personas: practice and theory. In *Proceedings of the 2003 conference on Designing for user experiences*. 1–15.
- [133] Katharina Reinecke and Krzysztof Z. Gajos. 2015. LabinthWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (Vancouver, BC, Canada) (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1364–1378. <https://doi.org/10.1145/2675133.2675246>
- [134] Katharina Reinecke, Tom Yeh, Luke Miratrix, Rahmatri Mardiko, Yuechen Zhao, Jenny Liu, and Krzysztof Z. Gajos. 2013. Predicting users’ first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) (CHI '13). Association for Computing Machinery, New York, NY, USA, 2049–2058. <https://doi.org/10.1145/2470654.2481281>
- [135] Anna Rogers, Niranjan Balasubramanian, Leon Derczynski, Jesse Dodge, Alexander Koller, Sasha Luccioni, Maarten Sap, Roy Schwartz, Noah A. Smith, and Emma Strubell. 2023. Closed AI Models Make Bad Baselines. <https://hackingsemantics.xyz/2023/closed-baselines/>
- [136] Kavous Salehzadeh Niksirat, Lahari Goswami, Pooja S. B. Rao, James Tyler, Alessandro Silacci, Sadiq Aliyu, Annika Aebli, Chat Wacharamanatham, and Mauro Cherubini. 2023. Changes in Research Ethics, Openness, and Transparency in Empirical Studies between CHI 2017 and CHI 2022. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 505, 23 pages. <https://doi.org/10.1145/3544548.3580848>
- [137] Joni Salminen, Chang Liu, Wenjing Pian, Jianxing Chi, Essi Häyhänen, and Bernard J Jansen. 2024. Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 510, 20 pages. <https://doi.org/10.1145/3613904.3642036>
- [138] Mark A Schmuckler. 2001. What is ecological validity? A dimensional analysis. *Infancy* 2, 4 (2001), 419–436.
- [139] 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).
- [140] Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs

- in group ideation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [141] Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S Bernstein. 2024. Rehearsal: Simulating conflict to teach conflict resolution. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.
- [142] Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating Self-Guided Mental Health Interventions Through Human-Language Model Interaction: A Case Study of Cognitive Restructuring. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 700, 29 pages. <https://doi.org/10.1145/3613904.3642761>
- [143] Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [144] Hong Shen, Tianshi Li, Toby Jia-Jun Li, Joon Sung Park, and Diyi Yang. 2023. Shaping the emerging norms of using large language models in social computing research. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 569–571.
- [145] Donghoo Shin, Lucy Lu Wang, and Gary Hsieh. 2024. From Paper to Card: Transforming Design Implications with Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.
- [146] Jessie J Smith, Saleema Amershi, Solon Barocas, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Real ml: Recognizing, exploring, and articulating limitations of machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 587–597.
- [147] Evropi Stefanidi, Marit Bentvelzen, Pawel W. Woźniak, Thomas Kosch, Mikolaj P. Woźniak, Thomas Mildner, Stefan Schneegass, Heiko Müller, and Jasmin Niess. 2023. Literature Reviews in HCI: A Review of Reviews. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 509, 24 pages. <https://doi.org/10.1145/3544548.3581332>
- [148] Konstantin R. Strömel, Stanislas Henry, Tim Johansson, Jasmin Niess, and Pawel W. Woźniak. 2024. Narrating Fitness: Leveraging Large Language Models for Reflective Fitness Tracker Data Interpretation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 646, 16 pages. <https://doi.org/10.1145/3613904.3642032>
- [149] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1039, 19 pages. <https://doi.org/10.1145/3613904.3642754>
- [150] Jiao Sun, Tongshuang Wu, Yue Jiang, Ronil Awalegaonkar, Xi Victoria Lin, and Diyi Yang. 2022. Pretty princess vs. successful leader: Gender roles in greeting card messages. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [151] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. 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, 1609–1621. <https://doi.org/10.1145/3630106.3658992>
- [152] Maryam Taeb, Amanda Swearning, Eldon Schoop, Ruijia Cheng, Yue Jiang, and Jeffrey Nichols. 2024. AXNav: Replying Accessibility Tests from Natural Language. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 962, 16 pages. <https://doi.org/10.1145/3613904.3642777>
- [153] Mei Tan and Hari Subramonyam. 2024. More than Model Documentation: Uncovering Teachers' Bespoke Information Needs for Informed Classroom Integration of ChatGPT. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 269, 19 pages. <https://doi.org/10.1145/3613904.3642592>
- [154] Yilin Tang, Liqing Chen, Ziyu Chen, Wenkai Chen, Yu Cai, Yao Du, Fan Yang, and Lingyun Sun. 2024. EmoEdu: Applying Generative Artificial Intelligence to Emotional Learning for Children with High-Function Autism. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1001, 20 pages. <https://doi.org/10.1145/3613904.3642899>
- [155] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2024. Debate Chatbots to Facilitate Critical Thinking on YouTube: Social Identity and Conversational Style Make A Difference. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 805, 24 pages. <https://doi.org/10.1145/3613904.3642513>
- [156] Nina Tran, Richard E Ladner, and Danielle Bragg. 2023. US Deaf Community Perspectives on Automatic Sign Language Translation. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–7.
- [157] Stephanie Valencia, Richard Cave, Krystal Kallarackal, Katie Seaver, Michael Terry, and Shaun K Kane. 2023. "The less I type, the better": How AI Language Models can Enhance or Impede Communication for AAC Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [158] Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2024. Large language models cannot replace human participants because they cannot portray identity groups. *arXiv preprint arXiv:2402.01908* (2024).
- [159] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling Conversational Interaction with Mobile UI using Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 432, 17 pages. <https://doi.org/10.1145/3544548.3580895>
- [160] Jiayao Wang, Haolong Hu, Zuyuan Wang, Song Yan, Youyu Sheng, and Dengbo He. 2024. Evaluating Large Language Models on Academic Literature Understanding and Review: An Empirical Study among Early-stage Scholars. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 12, 18 pages. <https://doi.org/10.1145/3613904.3641917>
- [161] Sitong Wang, Savvas Petridis, Taeahn Kwon, Xiaojuan Ma, and Lydia B Chilton. 2023. PopBlends: Strategies for conceptual blending with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [162] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 303, 21 pages. <https://doi.org/10.1145/3613904.3641960>
- [163] Zijie J Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–40.
- [164] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).
- [165] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.
- [166] Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. 2024. "As an AI language model, I cannot": Investigating LLM Denials of User Requests. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 979, 14 pages. <https://doi.org/10.1145/3613904.3642135>
- [167] Jules White, Sam Hays, Quichen Fu, Jesse Spencer-Smith, and Douglas C. Schmidt. 2024. *ChatGPT Prompt Patterns for Improving Code Quality, Refactoring, Requirements Elicitation, and Software Design*. Springer Nature Switzerland, Cham, 71–108. [https://doi.org/10.1007/978-3-031-55642-5\\_4](https://doi.org/10.1007/978-3-031-55642-5_4)
- [168] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research contributions in human-computer interaction. *Interactions* 23, 3 (apr 2016), 38–44. <https://doi.org/10.1145/2907069>
- [169] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 385, 22 pages. <https://doi.org/10.1145/3491102.3517582>
- [170] Dirk U Wulff, Zak Hussain, and Rui Mata. 2024. The Behavioral and Social Sciences Need Open LLMs. <https://doi.org/10.31219/osf.io/ybvz5>
- [171] Wei Xiang, Hanfei Zhu, Suqi Lou, Xinli Chen, Zhenghua Pan, Yiping Jin, Shi Chen, and Lingyun Sun. 2024. SimUser: Generating Usability Feedback by Simulating Various Users Interacting with Mobile Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [172] Ziang Xiao, Wesley Hanwen Deng, Michelle S. Lam, Motahhare Eslami, Juho Kim, Mina Lee, and Q. Vera Liao. 2024. Human-Centered Evaluation and Auditing of Language Models. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI EA '24). Association for Computing Machinery, New York, NY, USA, Article 476, 6 pages. <https://doi.org/10.1145/3613905.3636302>
- [173] Anna Xyglou, Chee Siang Ang, Panote Siriara, Jonasz Piotr Kopecki, Alexandra Covaci, Eiman Kanjo, and Wan-Jou She. 2024. MindTalker: Navigating the Complexities of AI-Enhanced Social Engagement for People with Early-Stage Dementia. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

- [174] Litao Yan, Alyssa Hwang, Zhiyuan Wu, and Andrew Head. 2024. Ivie: Light-weight anchored explanations of just-generated code. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.
- [175] Qian Yang, Nikola Banovic, and John Zimmerman. 2018. Mapping Machine Learning Advances from HCI Research to Reveal Starting Places for Design Innovation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–11. <https://doi.org/10.1145/3173574.3173704>
- [176] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T Iqbal, and Jaime Teevan. 2019. Sketching nlp: A case study of exploring the right things to design with language intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [177] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing Biomedical Literature to Calibrate Clinicians' Trust in AI Decision Support Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 14, 14 pages. <https://doi.org/10.1145/3544548.3581393>
- [178] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.
- [179] Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho De Castro, Shruthi Bannur, Stephanie Hyland, et al. 2024. Multimodal healthcare AI: identifying and designing clinically relevant vision-language applications for radiology. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.
- [180] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. <https://doi.org/10.1145/3544548.3581388>
- [181] Chao Zhang, Xuechen Liu, Katherine Ziska, Soobin Jeon, Chi-Lin Yu, and Ying Xu. 2024. Mathemyths: leveraging large language models to teach mathematical language through Child-AI co-creative storytelling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–23.
- [182] Lotus Zhang, Abigale Stangl, Tanusree Sharma, Yu-Yun Tseng, Inan Xu, Danna Gurari, Yang Wang, and Leah Findlater. 2024. Designing Accessible Obfuscation Support for Blind Individuals' Visual Privacy Management. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.
- [183] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792* (2023).
- [184] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.
- [185] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
- [186] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [187] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics* 50, 1 (2024), 237–291.
- [188] Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. 2024. Memoro: Using Large Language Models to Realize a Concise Interface for Real-Time Memory Augmentation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.