Caught in the Cascade: Why LLM Auditing is Missing the Middle

Anna Neumann
neumann@rc-trust.ai
Research Center Trustworthy Data Science and Security
Germany

Jatinder Singh
University of Cambridge
United Kingdom
Research Center Trustworthy Data Science and Security
Germany

Abstract

Large language models have transformed generative AI development, with foundation models serving as building blocks for diverse applications. The resulting auditing landscape focuses either on technical evaluations of foundation model capabilities or domain-specific assessments of deployed applications. However, these approaches miss crucial 'middle layers' that transform user inputs and model outputs before they reach the foundation model. These transformations can significantly alter system behaviour in ways neither foundation model nor application-level audits can detect and occur through components such as memory functions, system prompts, knowledge bases, and safety guardrails. Additionally, the transformations are operating in a hierarchy, allowing them to override each other. While such transformations potentially reshape the original intent of prompts, their effects vary in both magnitude and consequence. In practice, multiple stakeholders influence these layers but lack comprehensive visibility into both individual and cumulative effects. No single entity maintains oversight across the collective impact, hampering efforts to evaluate and audit the system as a whole. This position paper identifies gaps in current auditing approaches and indicates technical and human-centered research directions for evaluating transformations through intermediate layers.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Social and professional topics** → **Socio-technical systems**; **Technology audits**.

Keywords

Artificial Intelligence, Foundation Models, Large Language Models, Audit, Human-Centered, Domain-Specific

1 Introduction: 'Centering the Middle'

Large language models (LLMs) have fundamentally transformed how we develop and deploy artificial intelligence (AI) systems. The field has consolidated around foundation models that offer general capabilities [12, 102, 123], replacing specialized task-specific systems. These foundation models now serve as the basis for a wide range of applications [33, 102, 120], each adapted to specific contexts and requirements.

HEAL@CHI '25, April 26, 2025, Yokohama, Japan

Between user input and these foundation models lie multiple intermediate layers that progressively transform information, including memory functions [77], system prompts [109], and reasoning processes [31, 80, 111], each interpreting and potentially modifying the original input/output. These transformations occur both explicitly through direct modifications and implicitly through layer interactions, creating a cascade where changes influence subsequent processing steps. This relationship is shown in Figure 1 for LLM-based systems.

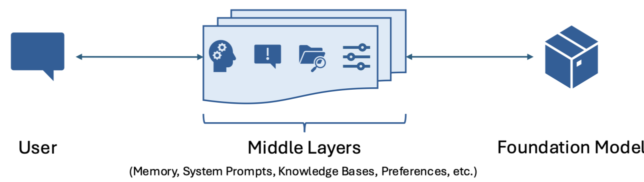

Figure 1: A user prompt can pass through multiple middle layers, such as memory functions, system prompts, knowledge bases, and user-specific preference data before reaching the 'black-box' foundation model and vice-versa.

1.1 Middle Layer Architectures

OpenAI's *'chain of command'*[79], provides a technical perspective on the hierarchical nature of these transformations. User prompts flow through increasingly powerful prompt categories: from user-derived guidelines and preferences, through developer-specified prompts, to platform-level modifications from OpenAI. Each category in this hierarchy can override prompts from previous levels, with platform-level modifications wielding the most power to reshape the input.

However, the chain of command represents just one dimension of these 'middle layers'. The full cascade encompasses a broader ecosystem: (possibly hidden) chain-of-thought or reasoning processes [31, 80, 111], memory functions [77] that retain and integrate past interactions, knowledge bases that supplement original inputs [36, 122, 125], tool use capabilities [44, 87, 110], further agentic functions [42], and both explicit and implicit safety guideline adaptations [34, 67]. Each element transforms not only how the foundation model processes the original query, but also how responses are generated, filtered, and presented back to the user—creating a bidirectional flow of transformations throughout the entire system.

1.2 Gaps in Auditing LLMs

In this position paper, we identify a critical gap in current approaches to auditing LLM-based systems: the lack of attention to how these middle layers transform user inputs and fundamentally shape system behavior. While existing auditing frameworks focus on either foundation models or domain-specific applications (see section 2), we argue that understanding the transformations occurring in these intermediate layers is essential for comprehensive system evaluation.

What makes addressing this gap particularly challenging is the potentially complex structures of these transformations. These interactions can overlap and override each other, creating transformation patterns where visibility is limited to one's own changes but not to how these interact with modifications from other layers. This fractured visibility creates a systemic evaluation challenge that current auditing approaches are ill-equipped to address.

1.3 Cascading Transformations: A Mental Health Example

To illustrate how these middle-layer transformations manifest in practice, consider a mental health support chatbot. An end-user writes 'I'm feeling overwhelmed.' The memory layer retrieves their recent messages about project deadlines and late-night work sessions. The therapeutic guidelines layer, seeing this work-related pattern, activates its workplace counseling protocols. The safety filter then combines these signals with built-in risk assessment rules, automatically escalating the case as potential burnout. A simple expression of momentary stress transforms into a high-priority workplace mental health case.[1]

This example reveals how traditional evaluation approaches fail to capture the full transformation process. Foundation model audits would reveal the model's general ability to recognize emotional distress, while application audits would show the final therapeutic interventions offered. However, neither approach captures how components like memory, knowledge bases, and safety filters function as crucial translation points within the system. Each component serves a specific function—memory systems retrieve contextual history, knowledge bases supply domain expertise, and safety filters implement protective boundaries. While these components can be evaluated individually, their sequential interactions create cascading transformations that standard evaluation methods fail to capture. A holistic system evaluation that examines these middle layers is necessary to understand the complete transformation process.

1.4 The Stakeholder Cascade

As foundation models are deployed across diverse applications [60], these middle layers grow in complexity and importance. Their cumulative effects create a cascade of transformations that becomes increasingly difficult to trace and evaluate across expanding data-driven AI supply chains[14, 27, 29]. This complexity is further compounded by the distributed responsibility across multiple stakeholders, each with different levels of system access and control [75].

The stakeholder cascade reflects the reality of modern LLM-based systems: developers implement general guardrails, deployers configure system prompts and tools, and end-users set preferences [4, 76]. In our mental health chatbot example, clinicians might implement therapeutic protocols based on established guidelines, while platform developers enforce content policies that may interact with those protocols in unpredictable ways. These elements function as crucial intermediaries in connecting foundation model capabilities with application behavior, ultimately shaping how individuals interact with and perceive the system.

This distributed 'responsibility' creates significant risks[29]: safety measures might be weakened when higher-priority rules override them, user preferences could be silently overridden by system-level constraints, and system behaviors may drift from their intended purposes as layers interact in unexpected ways. Moreover, when issues arise, the fragmented nature of these modifications makes it difficult to identify which layer or interaction caused the problem – creating ambiguous accountability and challenging troubleshooting.

These opaque transformative layers create challenges for system reliability, safety, and fairness across all LLM applications. While these issues are particularly critical in domains like healthcare, finance, and governance, they affect any context where users rely on LLMs for information processing or decision-making. Traditional auditing approaches, which focus on either foundation models or applications, fall short of addressing this multi-layered complexity, leaving a critical oversight gap in transparency [13] and accountability [29].

1.5 Our Contributions

This paper aims to brings attention to middle layers as a critical yet overlooked component in AI systems that requires dedicated consideration in system design, evaluation frameworks, and auditing methodologies. To this end, we make the following contributions:

- We analyze the current landscape of LLM-based AI auditing approaches and identify their limitations in addressing middle-layer transformations
- We define and characterize the cascading effects that occur as user inputs and outputs pass through multiple transformative layers, revealing their broader implications for transparency, accountability, and system performance
- We propose a research agenda spanning technical, human-centered, and integrated methodologies for evaluating middle-layer transformations in LLM-based systems

While we focus on systems built on or around LLMs in this paper, these principles and methodologies can be broadened to other model architectures and general purpose AI systems that employ similar layered transformation processes.

2 The Current LLM Auditing Landscape

Current algorithmic auditing approaches for generative language models center on two endpoints: foundation model capabilities and domain-specific applications. This 'dual-track' approach creates a significant gap in understanding how these systems operate in practice, particularly in transforming and processing information between these endpoints. While both approaches offer valuable insights, their separation limits our ability to comprehensively evaluate LLM-based systems.

[1] Such escalations may be appropriate and even desirable in certain contexts. Still, clarity about how the layers interact and shape outcomes helps us evaluate system behaviours.

This focus on endpoints leaves open questions about the transformative processes that occur between them – including the system's architectural layers and their impact. To understand these gaps, we first examine the current landscape of LLM-based system auditing.

2.1 Foundation Model Audits

Foundation model audits examine core model capabilities and limitations through both technical and broader conceptual analyses. Technical audits and evaluations assess fundamental behaviours [42, 49, 62, 80, 81], including task performance [3, 23, 44, 51, 106], bias patterns [45, 84, 95, 104], privacy measurements [83], and safety properties [94] through controlled experiments that measure outputs against predetermined benchmarks [6, 26, 88, 100].

Broader foundation model audits examine systemic implications [10, 113], specifically their development [55, 70, 82] and deployment [35, 65, 86, 116]. These studies analyze societal impacts [32, 33, 117], investigate ethical challenges [19, 43, 113], and evaluate approaches to algorithmic accountability [11, 29, 30, 90] and responsible AI development [16, 64, 118]. While valuable for understanding general implications, these analyses often remain disconnected from specific implementation contexts.

2.2 Domain-Specific Application Audits

Application-level audits evaluate deployed systems within specific real-world contexts, focusing on user experiences [25, 38, 39, 68], interaction patterns [66, 115], and domain-specific performance metrics [61, 71, 89, 96]. They often reveal how systems perform in practice, highlighting gaps between theoretical capabilities and actual performance. These assessments can generate targeted design recommendations [101, 108] that address particular application requirements and user needs [9, 46, 54, 92].

While these focused approaches provide valuable insights within their domains, they typically treat the foundation model as a black box [17], concentrating on final outputs and user interactions. In many cases, particularly where AI systems are integrated as one component in a larger supply chain [15, 28, 29], this black-box treatment is inherent to the system's architecture rather than a methodological choice. However, this opacity—whether by necessity or design—obscures the crucial transformations occurring between user input and system response.

2.3 Single-Layer Audits

Current research examines middle layers primarily in isolation, focusing on specific capabilities without capturing their interactions:

While foundation and application-level audits provide valuable insights, both typically operate with limited access to system internals [17]. Foundation model audits may examine certain technical components, e.g. model weights, architectures, and training data, but access remains restricted for both standard elements and the intermediate layers between foundation models and applications. This restricted visibility affects our understanding of how modifications and interactions occur through these intermediate layers. Current audits examine these middle layers in isolation, focusing on specific capabilities:

System messages [51, 52, 73, 74, 88] and instruction hierarchies form a 'chain of command' [79] that determines how models interpret and prioritize different directives [44, 109]. Research shows models can follow both individual simple [72, 109, 124] and complex instructions [47, 119], but questions remain about their effectiveness in real-world contexts where multiple directives may conflict [48, 50].

Evaluations of guideline following often focus on individual rules [34, 67, 91] or circumvention of these, i.e. jailbreaking [22, 100, 112], missing how guidelines interact with other system elements.

Research on generative AI highlights capabilities in agentic behaviors [2, 18, 21, 42, 63, 78], tool use [24, 93, 97, 105, 110], and knowledge base integration [37, 53, 56] – all representing middle layers that transform inputs between user requests and model outputs.

3 Missing the Middle: Middle-Layer Auditing Challenges

Current foundation model and application-specific auditing approaches fail to capture the critical middle layers where significant transformations occur, indicating a clear need for holistic evaluation methodologies. These methodologies must examine how multiple layers interact and collectively transform inputs and outputs throughout the system pipeline. Having established that current auditing approaches focus primarily on foundation models and applications while overlooking the crucial middle layers, we now examine the specific challenges these layers present for comprehensive system evaluation.

3.1 The Cascade Effect

At the core of this challenge is how middle layers modify inputs/outputs through sequential transformations, with each layer's changes affecting how subsequent layers process that input – what we term a **cascade**. Returning to our hypothetical mental health chatbot example: the chat history alters the user's initial input, therapeutic guidelines transform this modified input into an augmented context, and safety guidelines process this accumulated message before it reaches the foundation model and also process its output. Each step in this cascade not only transforms the input but shapes the context for all following transformations, potentially amplifying small initial changes into significant shifts in system behavior. These effects become more pronounced in complex systems where multiple layers interact hierarchically in various combinations.

It is important to note that we use *'cascade'* metaphorically to describe the cumulative flow of transformations [98], distinct from technical cascade models in machine learning or software engineering (e.g., cascading classifiers [1] or waterfall models [85]).

3.2 Opacity in AI Supply Chains

This complexity increases substantially when considering how LLM-based technologies are embedded within broader technological infrastructures. Consider a healthcare system where an LLM is integrated into clinical workflows: the LLM interacts with electronic health records that preprocess patient data, interfaces with clinical decision support systems that apply medical guidelines, and operates within institutional compliance frameworks that enforce privacy and security protocols. Each integration point introduces additional transformative layers that modify information flows. These complex integration environments, often referred to as AI supply chains [7, 8, 27, 29, 114], involve numerous stakeholders with differing levels of system access and control:

- Foundation model developers implement base model capabilities and general safeguards
- System integrators configure domain-specific policies and integration points
- Domain specialists define field-specific guidelines and terminology standards
- End-users adjust individual preferences and contextual parameters

Each stakeholder typically has visibility into only their portion of the system [58, 98]. A physician using an LLM-enhanced diagnostic tool may be unaware that their queries pass through privacy filters, medical terminology standardizers, and knowledge base augmenters before reaching the foundation model – each potentially altering the original question in significant ways. Technical documentation rarely explores these interaction effects, focusing instead on individual components rather than their combined behavior.

This limited visibility creates blind spots where transformations happen without any stakeholder having full oversight. The problem becomes even more complex when layers learn implicitly from user behaviour [79] rather than following explicit rules, making their effects harder to predict and audit.

The combination of *limited layer visibility* and *growing system complexity* points to the need for novel auditing methods that can capture both individual layer behaviors and their collective effects within and across AI systems [98, 99].

3.3 Key Research Challenges

Based on our analysis of middle-layer transformations and AI supply chains, we identify two fundamental challenges that any comprehensive research agenda must address [58, 98]:

- **Tracking transformations across interacting layers:** Existing approaches do not effectively track changes across multiple interacting components. While foundation model audits, layer-specific evaluations, and domain-focused assessments each provide valuable insights, they miss how transformations compound and interact in practice.
- **Managing increasing system complexity:** The growing sophistication of LLM supply chains introduces additional layers and interactions, making systematic analysis increasingly difficult [27] . Even with direct access to individual layers, current evaluation methods struggle to capture emergent effects from component interactions.

The combination of *limited layer visibility* and *growing system complexity* points to the need for novel auditing methods that can capture both individual layer behaviors and their collective effects across AI systems.

We further identify **five critical technical implications** of middle-layer processing that represent unique challenges for AI system design and evaluation. As shown in Table 1, these implications – semantic transformation, authority hierarchy, invisible processing, emergent behaviors, and scaling resistance – form a framework for understanding how cascading effects manifest throughout AI systems. Each implication highlights distinct challenges that current evaluation approaches fail to address. Together, they illustrate why middle layers demand specialized attention beyond traditional auditing methods and inform the research agenda in Section section 4.

Semantic Transformation	Each layer can fundamentally alter an input's meaning, intent, and structure beyond surface-level changes, potentially distorting user intent.
Authority Hierarchy	Layers operate in a hierarchy of authority, where higher-level transformations can override or nullify lower-level changes, obscuring accountability and making it difficult to trace the source of harmful outcomes.
Invisible Processing	These transformations occur invisibly to most stakeholders, preventing effective oversight of compounding changes throughout the system.
Emergent Behaviors	Layer interactions can create emergent behaviors that escape both foundation model and application-level audits but significantly impact system responses.
Scaling Resistance	Research shows that foundation models become less willing to modify their core behaviors and preferences as they scale up, suggesting fundamental limits to how much middle layers can reshape model responses.

Table 1: Critical Technical Implications of Middle-Layer Cascades

4 Mapping the Middle: A Research Agenda for Middle-Layer Auditing

To develop human-centered systems, we need better ways to understand and evaluate these transformative processes. System design must ensure that middle-layer modifications enhance rather than distort the user's intent. This focus becomes particularly important as foundation models expand into diverse applications, each requiring specific combinations of middle layers to bridge between model capabilities and application requirements.

The following section outlines a research agenda that addresses these challenges through technical, human-centered, and integrated approaches to understanding and evaluating middle-layer transformations.

4.1 Technical Research Directions

The technical implications we've identified suggest important research opportunities for the socio-technical research community:

- **Methods to track data flow** and transformations through middle layers
- **Fairness implications** of transformation hierarchies
- **Evaluation of system behavior** changes through different middle layer combinations

- **Assessment of safety mechanism robustness** across middle layer interactions

4.2 Human-Centered Investigation Approaches

The core challenge of human-centered middle-layer research is understanding the domain-specific contexts in which these systems operate. Different application domains fundamentally shape how these layers are configured and used – medical systems demand strict safety protocols, legal applications require precise terminology management, and educational tools need adaptable difficulty scaling. This context specificity means research must examine not just the technical aspects of middle layers, but how diverse human stakeholders interpret, configure, and interact with these systems in their particular domains.

The identified technical effects manifest in critical ways across deployed systems: sequential transformations can alter system behavior in ways stakeholders cannot predict [102, 121], accountability becomes fragmented across multiple parties [29, 69, 107], safety-critical applications face compounding risks [20, 41] and bias detection grows increasingly complex as discriminatory effects may emerge from cumulative interactions [40, 57, 59]. Importantly, these implications arise not from theoretical concerns but from the *everyday operation of deployed systems*, where middle layers interact with each other.

These real-world implications highlight opportunities for expanded research across disciplines:

- **Developer-focused studies** examining how technical teams implement domain-specific requirements through middle layer configurations
- **Domain expert investigations** exploring how professionals in different fields interact with and adapt middle layer configurations [5, 103]
- **Cross-domain analyses** comparing how different sectors handle similar challenges using middle layers
- **Longitudinal analyses** tracking how domain-specific requirements evolve and middle layer configurations adapt over time

4.3 Integration Approaches

Beyond isolated technical or human-centered methods, understanding these systems requires approaches that bridge multiple perspectives and methodologies. Several integrated approaches offer potential:

- **Participatory audit frameworks** combining stakeholder expertise with technical analysis
- **Mixed-method studies** combining layer tracking with user impact assessment
- **Cross-disciplinary evaluation** frameworks incorporating both system behaviour and organizational context
- **Combined analysis** of technical fairness metrics and stakeholder fairness perceptions
- **Implication and risk analysis** of middle layer auditing; looking at risks of transparency – e.g., exposing safety guardrails might ease adversarial attacks.

These integrated approaches could reveal how intermediate layers shape system behaviour across different contexts and stakeholder groups. By combining technical analysis with human perspectives, they offer directions for examining the full scope of layer interactions in LLM-based systems.

5 Conclusion

Current LLM-based auditing approaches miss crucial transformative processes occurring in middle layers – from system prompts to memory functions. These layers fundamentally shape system behaviour through complex interactions that neither foundation model nor application-level audits can fully capture.

This paper identifies the need to examine how these layers modify and interact with various inputs/outputs, highlighting a gap in existing audit frameworks. We outline potential technical, human-centered, and integrated approaches as starting points for researchers to examine these layers, their interactions, and their impacts.

Future research in this direction could enable more holistic interventions to improve safety and fairness, provide insights into how different stakeholders shape system behaviour across data-driven AI supply chains, and reveal how components affect user experiences. As LLM-based systems grow more complex, understanding middle-layer transformations becomes increasingly important for effective system evaluation and responsible AI development. While our focus has been on LLM-based systems, the principles and methodologies we propose can be extended to other general-purpose AI architectures that employ similar layered transformation processes (in practice, many AI service-based infrastructures), including multimodal systems, autonomous agents, and future AI architectures. We call on human-AI interaction researchers across disciplines to prioritize this critical area of investigation and develop robust frameworks for middle-layer auditing.

References

[1] Ethem Alpaydin and Cenk Kaynak. 1998. Cascading classifiers. *Kybernetika* 34, 4 (1998), 369–374.

[2] Anthropic. 2024. Building effective agents. https://www.anthropic.com/research/building-effective-agents

[3] Anthropic. 2024. Claude 3 model card. https://docs.anthropic.com/en/docs/resources/model-card

[4] Anthropic. 2024. Understanding Claude's Personalization Features | Anthropic Help Center. https://support.anthropic.com/en/articles/10185728-understanding-claude-s-personalization-features

[5] Ian Arawjo, Priyan Vaithilingam, Martin Wattenberg, and Elena Glassman. 2023. ChainForge: An open-source visual programming environment for prompt engineering. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 4, 3 pages. https://doi.org/10.1145/3586182.3616660

[6] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023. Benchmarking Foundation Models with Language-Model-as-an-Examiner. *Advances in Neural Information Processing Systems* 36 (Dec. 2023), 78142–78167. https://proceedings.neurips.cc/paper_files/paper/2023/hash/f64e55d03e2fe61aa4114e49cb654acb-Abstract-Datasets_and_Benchmarks.html

[7] Agathe Balayn, Yulu Pi, David Gray Widder, Kars Alfrink, Mireia Yurrita, Sohini Upadhyay, Naveena Karusala, Henrietta Lyons, Cagatay Turkay, Christelle Tessono, Blair Attard-Frost, and Ujwal Gadiraju. 2024. From Stem to Stern: Contestability Along AI Value Chains. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. ACM, San Jose Costa Rica, 720–723. https://doi.org/10.1145/3678884.3681831

[8] Agathe Balayn, Yulu Pi, David Gray Widder, Kars Alfrink, Mireia Yurrita, Sohini Upadhyay, Naveena Karusala, Henrietta Lyons, Cagatay Turkay, Christelle Tessono, Blair Attard-Frost, and Ujwal Gadiraju. 2024. From Stem to Stern: Contestability Along AI Value Chains. In *Companion Publication of the 2024*

Conference on Computer-Supported Cooperative Work and Social Computing (San Jose, Costa Rica) *(CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 720–723. https://doi.org/10.1145/3678884.3681831

[9] Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2021. Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 116–128. https://doi.org/10.1145/3442188.3445875

[10] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[11] Daniel James Bogiatzis-Gibbons. 2024. Beyond Individual Accountability: (Re-)Asserting Democratic Control of AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 74–84. https://doi.org/10.1145/3630106.3658541

[12] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/arXiv.2108.07258 arXiv:2108.07258 [cs].

[13] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. The foundation model transparency index. *arXiv preprint arXiv:2310.12941* (2023).

[14] Alexandra Brintrup, George Baryannis, Ashutosh Tiwari, Svetan Ratchev, Giovanna Martinez-Arellano, and Jatinder Singh. 2023. Trustworthy, responsible, ethical AI in manufacturing and supply chains: synthesis and emerging research questions. https://doi.org/10.48550/arXiv.2305.11581 arXiv:2305.11581 [cs].

[15] Alexandra Brintrup, George Baryannis, Ashutosh Tiwari, Svetan Ratchev, Giovanna Martinez-Arellano, and Jatinder Singh. 2023. Trustworthy, responsible, ethical AI in manufacturing and supply chains: synthesis and emerging research questions. arXiv:2305.11581 [cs.AI] https://arxiv.org/abs/2305.11581

[16] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, J. B. Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. https://doi.org/10.48550/arXiv.2004.07213 arXiv:2004.07213 [cs].

[17] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 2254–2272. https://doi.org/10.1145/3630106.3659037

[18] Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. 2024. Visibility into AI Agents. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 958–973. https://doi.org/10.1145/3630106.3658948

[19] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 651–666. https://doi.org/10.1145/3593013.3594033

[20] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 651–666. https://doi.org/10.1145/3593013.3594033

[21] Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. 2024. MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. https://doi.org/10.48550/arXiv.2410.07095 arXiv:2410.07095 [cs].

[22] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2024. Jailbreaking Black Box Large Language Models in Twenty Queries. https://doi.org/10.48550/arXiv.2310.08419 arXiv:2310.08419 [cs].

[23] Shijie Chen, Yu Zhang, and Qiang Yang. 2024. Multi-Task Learning in Natural Language Processing: An Overview. *ACM Comput. Surv.* 56, 12, Article 295 (July 2024), 32 pages. https://doi.org/10.1145/3663363

[24] Zhi-Yuan Chen, Shiqi Shen, Guangyao Shen, Gong Zhi, Xu Chen, and Yankai Lin. 2024. Towards Tool Use Alignment of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 1382–1400. https://doi.org/10.18653/v1/2024.emnlp-main.82

[25] Inyoung Cheong, King Xia, KJ Kevin Feng, Quan Ze Chen, and Amy X Zhang. 2024. (A) I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2454–2469.

[26] François Chollet. 2019. On the Measure of Intelligence. https://doi.org/10.48550/arXiv.1911.01547 arXiv:1911.01547 [cs].

[27] Jennifer Cobbe and Jatinder Singh. 2021. Artificial intelligence as a service: Legal responsibilities, liabilities, and policy challenges. *Computer Law & Security Review* 42 (2021), 105573. https://www.sciencedirect.com/science/article/pii/S0267364921000467 Publisher: Elsevier.

[28] Jennifer Cobbe and Jatinder Singh. 2021. Artificial intelligence as a service: Legal responsibilities, liabilities, and policy challenges. *Computer Law Security Review* 42 (2021), 105573. https://doi.org/10.1016/j.clsr.2021.105573

[29] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding accountability in algorithmic supply chains. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1186–1197. https://doi.org/10.1145/3593013.3594073

[30] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1571–1583. https://doi.org/10.1145/3531146.3533213

[31] DeepSeek-AI et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. https://doi.org/10.48550/arXiv.2501.12948 arXiv:2501.12948 [cs].

[32] Nathalie Diberardino, Clair Baleshta, and Luke Stark. 2024. Algorithmic Harms and Algorithmic Wrongs. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1725–1732. https://doi.org/10.1145/3630106.3659001

[33] Andrés Domínguez Hernández, Shyam Krishna, Antonella Maia Perini, Michael Katell, Sj Bennett, Ann Borda, Youmna Hashem, Semeli Hadjiloizou, Sabeehah Mahomed, Smera Jayadeva, Mhairi Aitken, and David Leslie. 2024. Mapping the individual, social and biospheric impacts of Foundation Models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 776–796. https://doi.org/10.1145/3630106.3658939

[34] Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, Saddek Bensalem, and Xiaowei Huang.

2024. Safeguarding Large Language Models: A Survey. arXiv:2406.02622 [cs.CR] https://arxiv.org/abs/2406.02622

[35] Sabri Eyuboglu, Karan Goel, Arjun Desai, Lingjiao Chen, Mathew Monfort, Chris Ré, and James Zou. 2024. Model ChangeLists: Characterizing Updates to ML Models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 2432–2453. https://doi.org/10.1145/3630106.3659047

[36] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, Barcelona Spain, 6491–6501. https://doi.org/10.1145/3637528.3671470

[37] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. arXiv:2405.06211 [cs.CL] https://arxiv.org/abs/2405.06211

[38] Anjalie Field, Amanda Coston, Nupoor Gandhi, Alexandra Chouldechova, Emily Putnam-Hornstein, David Steier, and Yulia Tsvetkov. 2023. Examining risks of racial biases in NLP tools for child protective services. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1479–1492. https://doi.org/10.1145/3593013.3594094

[39] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 205–216. https://doi.org/10.1145/3593013.3593989

[40] Kate Glazko, Yusuf Mohammed, Ben Kosa, Venkatesh Potluri, and Jennifer Mankoff. 2024. Identifying and Improving Disability Bias in GPT-Based Resume Screening. 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, 687–700. https://doi.org/10.1145/3630106.3658933

[41] Delaram Golpayegani, Harshvardhan J. Pandit, and Dave Lewis. 2023. To Be High-Risk, or Not To Be—Semantic Specifications and Implications of the AI Act's High-Risk AI Applications and Harmonised Standards. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 905–915. https://doi.org/10.1145/3593013.3594050

[42] Google. 2024. Introducing Gemini 2.0: our new AI model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/

[43] Mélanie Gornet, Simon Delarue, Maria Boritchev, and Tiphaine Viard. 2024. Mapping AI ethics: a meso-scale analysis of its charters and manifestos. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 127–140. https://doi.org/10.1145/3630106.3658545

[44] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, et al. 2024. The Llama 3 Herd of Models. https://doi.org/10.48550/arXiv.2407.21783 arXiv:2407.21783 [cs].

[45] Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. https://doi.org/10.48550/arXiv.2311.04892 arXiv:2311.04892 [cs].

[46] Jamie Hancock, Ruoyun Hui, Jatinder Singh, and Anjali Mazumder. 2024. Trouble at Sea: Data and digital technology challenges for maritime human rights concerns. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 988–1001. https://doi.org/10.1145/3630106.3658950

[47] Qianyu He, Jie Zeng, Qianxi He, Jiaqing Liang, and Yanghua Xiao. 2024. From Complex to Simple: Enhancing Multi-Constraint Complex Instruction Following Ability of Large Language Models. https://doi.org/10.48550/arXiv.2404.15846 arXiv:2404.15846 [cs].

[48] Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024. Can Large Language Models Understand Real-World Complex Instructions? *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 16 (March 2024), 18188–18196. https://doi.org/10.1609/aaai.v38i16.29777 Number: 16.

[49] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. https://doi.org/10.48550/arXiv.2009.03300 arXiv:2009.03300 [cs].

[50] Yerin Hwang, Yongil Kim, Jahyun Koo, Taegwan Kang, Hyunkyung Bae, and Kyomin Jung. 2025. LLMs can be easily Confused by Instructional Distractions. https://doi.org/10.48550/arXiv.2502.04362 arXiv:2502.04362 [cs].

[51] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. https://doi.org/10.48550/arXiv.2310.06825 arXiv:2310.06825 [cs].

[52] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. https://doi.org/10.48550/arXiv.2305.02547 arXiv:2305.02547 [cs].

[53] Erik Jones, Anca Dragan, Aditi Raghunathan, and Jacob Steinhardt. 2023. Automatically Auditing Large Language Models via Discrete Optimization. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 15307–15329. https://proceedings.mlr.press/v202/jones23a.html ISSN: 2640-3498.

[54] Chowon Kang, Yoonseo Choi, Yongjae Sohn, Hyunseung Lim, and Hwajung Hong. 2024. Beyond Swipes and Scores: Investigating Practices, Challenges and User-Centered Values in Online Dating Algorithms. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 486 (Nov. 2024), 30 pages. https://doi.org/10.1145/3687025

[55] Bhargav Kumar Konidena, Jesu Narkarunai Arasu Malaiyappan, and Anish Tadimarri. 2024. Ethical Considerations in the Development and Deployment of AI Systems. *European Journal of Technology* 8, 2 (March 2024), 41–53. https://doi.org/10.47672/ejt.1890 Number: 2.

[56] Angelie Kraft and Eloïse Soulier. 2024. Knowledge-Enhanced Language Models Are Not Bias-Proof: Situated Knowledge and Epistemic Injustice in AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1433–1445. https://doi.org/10.1145/3630106.3658981

[57] Angelie Kraft and Eloïse Soulier. 2024. Knowledge-Enhanced Language Models Are Not Bias-Proof: Situated Knowledge and Epistemic Injustice in AI. 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, 1433–1445. https://doi.org/10.1145/3630106.3658981

[58] Joshua A. Kroll. 2021. Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. ACM, 758–771. https://doi.org/10.1145/3442188.3445937

[59] Kornel Lewicki, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2023. Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service". In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3544548.3581463

[60] Bo Li, Dongseong Hwang, Zhouyuan Huo, Junwen Bai, Guru Prakash, Tara N. Sainath, Khe Chai Sim, Yu Zhang, Wei Han, Trevor Strohman, and Francoise Beaufays. 2023. Efficient Domain Adaptation for Speech Foundation Models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096330

[61] Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, et al. 2024. LegalAgentBench: Evaluating LLM Agents in Legal Domain. *arXiv preprint arXiv:2412.17259* (2024).

[62] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. https://doi.org/10.48550/arXiv.2211.09110 arXiv:2211.09110 [cs].

[63] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023. AgentBench: Evaluating

LLMs as Agents. https://doi.org/10.48550/arXiv.2308.03688 arXiv:2308.03688 [cs].
[64] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2024. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv:2308.05374 [cs.AI] https://arxiv.org/abs/2308.05374
[65] 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*. ACM, Rio de Janeiro Brazil, 85–99. https://doi.org/10.1145/3630106.3658542
[66] Michael Madaio, Shivani Kapania, Rida Qadri, Ding Wang, Andrew Zaldivar, Remi Denton, and Lauren Wilcox. 2024. Learning about Responsible AI On-The-Job: Learning Pathways, Orientations, and Aspirations. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1544–1558. https://doi.org/10.1145/3630106.3658988
[67] Yaaseen Mahomed, Charlie M. Crawford, Sanjana Gautam, Sorelle A. Friedler, and Danaë Metaxa. 2024. Auditing GPT's Content Moderation Guardrails: Can ChatGPT Write Your Favorite TV Show?. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 660–686. https://doi.org/10.1145/3630106.3658932
[68] Samuel Mayworm, Kendra Albert, and Oliver L. Haimson. 2024. Misgendered During Moderation: How Transgender Bodies Make Visible Cisnormative Content Moderation Policies and Enforcement in a Meta Oversight Board Case. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 301–312. https://doi.org/10.1145/3630106.3658907
[69] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 735–746. https://doi.org/10.1145/3442188.3445935
[70] Margaret Mitchell, Avijit Ghosh, Alexandra Sasha Luccioni, and Giada Pistilli. 2025. Fully Autonomous AI Agents Should Not be Developed. arXiv:2502.02649 [cs.AI] https://arxiv.org/abs/2502.02649
[71] Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. WorldBench: Quantifying Geographic Disparities in LLM Factual Recall. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1211–1228.
[72] Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Basel Alomair, Dan Hendrycks, and David Wagner. 2024. Can LLMs Follow Simple Rules? https://doi.org/10.48550/arXiv.2311.04235 arXiv:2311.04235 [cs].
[73] Norman Mu, Jonathan Lu, Michael Lavery, and David Wagner. 2024. A Closer Look at System Message Robustness. https://openreview.net/forum?id=YZqDyqYwFf
[74] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive Learning from Complex Explanation Traces of GPT-4. https://doi.org/10.48550/arXiv.2306.02707 arXiv:2306.02707 [cs].
[75] Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and Engineering Ethics* 2, 1 (March 1996), 25–42. https://doi.org/10.1007/BF02639315
[76] OpenAI. 2024. Custom instructions for ChatGPT. https://openai.com/index/custom-instructions-for-chatgpt/
[77] OpenAI. 2024. Memory and new controls for ChatGPT. https://openai.com/index/memory-and-new-controls-for-chatgpt/
[78] OpenAI. 2025. Computer-Using Agent. https://openai.com/index/computer-using-agent/
[79] OpenAI. 2025. OpenAI Model Spec. https://model-spec.openai.com/2025-02-12.html
[80] OpenAI. 2025. OpenAI o1 System Card. https://openai.com/index/openai-o1-system-card/
[81] OpenAI. 2025. OpenAI o3-mini System Card. https://openai.com/index/o3-mini-system-card/
[82] Will Orr and Edward B. Kang. 2024. AI as a Sport: On the Competitive Epistemologies of Benchmarking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1875–1884. https://doi.org/10.1145/3630106.3659012
[83] Ashwinee Panda, Xinyu Tang, Milad Nasr, Christopher A. Choquette-Choo, and Prateek Mittal. 2024. Privacy Auditing of Large Language Models. https://openreview.net/forum?id=6mVZUh4kkY
[84] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A Hand-Built Bias Benchmark for Question Answering. https://doi.org/10.48550/arXiv.2110.08193 arXiv:2110.08193 [cs].
[85] Kai Petersen, Claes Wohlin, and Dejan Baca. 2009. The waterfall model in large-scale development. In *Product-Focused Software Process Improvement: 10th International Conference, PROFES 2009, Oulu, Finland, June 15-17, 2009. Proceedings 10*. Springer, 386–400.
[86] Aisyah Putri and Minh Quang Tran. 2023. Global Perspectives on AI Deployment: Cultural, Ethical, and Operational Dimensions. *Journal of Computational Social Dynamics* 8, 11 (Nov. 2023), 26–33. https://vectoral.org/index.php/JCSD/article/view/87
[87] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Guoliang Li, Zhiyuan Liu, and Maosong Sun. 2024. Tool Learning with Foundation Models. *ACM Comput. Surv.* 57, 4, Article 101 (Dec. 2024), 40 pages. https://doi.org/10.1145/3704435
[88] Yanzhao Qin, Tao Zhang, Tao Zhang, Yanjun Shen, Wenjing Luo, Haoze Sun, Yan Zhang, Yujing Qiao, Weipeng Chen, Zenan Zhou, Wentao Zhang, and Bin Cui. 2024. SysBench: Can Large Language Models Follow System Messages? https://doi.org/10.48550/arXiv.2408.10943 arXiv:2408.10943 [cs].
[89] Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J. Topol. 2024. LLM-based agentic systems in medicine and healthcare. *Nature Machine Intelligence* 6, 12 (Dec. 2024), 1418–1420. https://doi.org/10.1038/s42256-024-00944-1 Publisher: Nature Publishing Group.
[90] Divya Ramesh, Vaishnav Kameswaran, Ding Wang, and Nithya Sambasivan. 2022. How Platform-User Power Relations Shape Algorithmic Accountability: A Case Study of Instant Loan Platforms and Financially Stressed Users in India. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1917–1928. https://doi.org/10.1145/3531146.3533237
[91] Qingyu Ren, Jie Zeng, Qianyu He, Jiaqing Liang, Yanghua Xiao, Weikang Zhou, Zeye Sun, and Fei Yu. 2025. Step-by-Step Mastery: Enhancing Soft Constraint Following Ability of Large Language Models. arXiv:2501.04945 [cs.CL] https://arxiv.org/abs/2501.04945
[92] Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. 2024. Escalation Risks from Language Models in Military and Diplomatic Decision-Making. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 836–898. https://doi.org/10.1145/3630106.3658942
[93] Jingqing Ruan, YiHong Chen, Bin Zhang, Zhiwei Xu, Tianpeng Bao, Du Guo Qing, Shi Shiwei, Hangyu Mao, Ziyue Li, Xingyu Zeng, and Rui Zhao. 2023. TPTU: Task Planning and Tool Usage of Large Language Model-based AI Agents. https://openreview.net/forum?id=GrkgKtOjaH
[94] Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. https://doi.org/10.48550/arXiv.2308.01263 arXiv:2308.01263 [cs].
[95] Alejandro Salinas, Amit Haim, and Julian Nyarko. 2025. What's in a Name? Auditing Large Language Models for Race and Gender Bias. https://doi.org/10.48550/arXiv.2402.14875 arXiv:2402.14875 [cs].
[96] Bianca Giulia Sarah Schor, Emma Kallina, Jatinder Singh, and Alan Blackwell. 2024. Meaningful Transparency for Clinicians: Operationalising HCXAI Research with Gynaecologists. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1268–1281. https://doi.org/10.1145/3630106.3658971
[97] Zhengliang Shi, Shen Gao, Xiuyi Chen, Yue Feng, Lingyong Yan, Haibo Shi, Dawei Yin, Zhumin Chen, Suzan Verberne, and Zhaochun Ren. 2024. Chain of Tools: Large Language Model is an Automatic Multi-tool Learner. https://doi.org/10.48550/arXiv.2405.16533 arXiv:2405.16533 [cs].
[98] Jatinder Singh, Jennifer Cobbe, and Chris Norval. 2019. Decision Provenance: Harnessing Data Flow for Accountable Systems. *IEEE Access* 7 (2019), 6562–6574. https://doi.org/10.1109/ACCESS.2018.2887201
[99] Jatinder Singh, Ian Walden, Jon A. Crowcroft, and Jean Bacon. 2016. Responsibility & Machine Learning: Part of a Process. *Social Science Research Network* (2016). https://api.semanticscholar.org/CorpusID:168317965
[100] Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. A StrongREJECT for Empty Jailbreaks. https://doi.org/10.48550/arXiv.2402.10260 arXiv:2402.10260 [cs].
[101] Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer. 2024. On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 2495–2507. https://doi.org/10.1145/3630106.3659051
[102] Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1609–1621. https://doi.org/10.1145/3630106.3658992
[103] Annalisa Szymanski, Simret Araya Gebreegziabher, Oghenemaro Anuyah, Ronald A. Metoyer, and Toby Jia-Jun Li. 2024. Comparing Criteria Development Across Domain Experts, Lay Users, and Models in Large Language Model Evaluation. arXiv:2410.02054 [cs.HC] https://arxiv.org/abs/2410.02054

[104] Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and Mitigating Discrimination in Language Model Decisions. https://doi.org/10.48550/arXiv.2312.03689 arXiv:2312.03689 [cs].

[105] Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. 2023. ToolAlpaca: Generalized Tool Learning for Language Models with 3000 Simulated Cases. https://doi.org/10.48550/arXiv.2306.05301 arXiv:2306.05301 [cs].

[106] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. https://doi.org/10.48550/arXiv.2307.09288 arXiv:2307.09288 [cs].

[107] Rama Adithya Varanasi and Nitesh Goyal. 2023. "It is currently hodgepodge": Examining AI/ML Practitioners' Challenges during Co-production of Responsible AI Values. 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 251, 17 pages. https://doi.org/10.1145/3544548.3580903

[108] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. 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, 1–14. https://doi.org/10.1145/3173574.3174014

[109] Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. 2024. The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions. https://doi.org/10.48550/arXiv.2404.13208 arXiv:2404.13208 [cs] version: 1.

[110] Zhiruo Wang, Zhoujun Cheng, Hao Zhu, Daniel Fried, and Graham Neubig. 2024. What Are Tools Anyway? A Survey from the Language Model Perspective. https://doi.org/10.48550/arXiv.2403.15452 arXiv:2403.15452 [cs].

[111] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. https://doi.org/10.48550/arXiv.2201.11903 arXiv:2201.11903 [cs].

[112] Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, and Yisen Wang. 2024. Jailbreak and Guard Aligned Language Models with Only Few In-Context Demonstrations. https://doi.org/10.48550/arXiv.2310.06387 arXiv:2310.06387 [cs].

[113] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. https://doi.org/10.48550/arXiv.2112.04359 arXiv:2112.04359 [cs].

[114] David Gray Widder and Dawn Nafus. 2023. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society* 10, 1 (2023), 20539517231177620. https://doi.org/10.1177/20539517231177620 arXiv:https://doi.org/10.1177/20539517231177620

[115] David Gray Widder, Derrick Zhen, Laura Dabbish, and James Herbsleb. 2023. It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them?. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 467–479. https://doi.org/10.1145/3593013.3594012

[116] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. 2022. Sustainable AI: Environmental Implications, Challenges and Opportunities. In *Proceedings of Machine Learning and Systems*, D. Marculescu, Y. Chi, and C. Wu (Eds.), Vol. 4. 795–813. https://proceedings.mlsys.org/paper_files/paper/2022/file/462211f67c7d858f663355eff93b745e-Paper.pdf

[117] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. 2022. Sustainable AI: Environmental Implications, Challenges and Opportunities. *Proceedings of Machine Learning and Systems* 4 (April 2022), 795–813. https://proceedings.mlsys.org/paper_files/paper/2022/hash/462211f67c7d858f663355eff93b745e-Abstract.html

[118] Boming Xia, Qinghua Lu, Liming Zhu, Sung Une Lee, Yue Liu, and Zhenchang Xing. 2024. Towards a Responsible AI Metrics Catalogue: A Collection of Metrics for AI Accountability. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI* (Lisbon, Portugal) *(CAIN '24)*. Association for Computing Machinery, New York, NY, USA, 100–111. https://doi.org/10.1145/3644815.3644959

[119] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. WizardLM: Empowering Large Language Models to Follow Complex Instructions. https://doi.org/10.48550/arXiv.2304.12244 arXiv:2304.12244 [cs].

[120] Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. 2023. Foundation Models for Decision Making: Problems, Methods, and Opportunities. https://doi.org/10.48550/arXiv.2303.04129 arXiv:2303.04129 [cs].

[121] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. 2022. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 535–563. https://doi.org/10.1145/3531146.3533118

[122] Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. ∞Bench: Extending Long Context Evaluation Beyond 100K Tokens. https://doi.org/10.48550/arXiv.2402.13718 arXiv:2402.13718 [cs].

[123] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. 2024. A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. *International Journal of Machine Learning and Cybernetics* (Nov. 2024). https://doi.org/10.1007/s13042-024-02443-6

[124] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-Following Evaluation for Large Language Models. https://doi.org/10.48550/arXiv.2311.07911 arXiv:2311.07911 [cs].

[125] Yuqi Zhu, Shuofei Qiao, Yixin Ou, Shumin Deng, Ningyu Zhang, Shiwei Lyu, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024. Knowagent: Knowledge-augmented planning for llm-based agents. *arXiv preprint arXiv:2403.03101* (2024).

Received 24 February 2025; revised 14 April 2025