A collection of principles for guiding and evaluating large language models

Konstantin Hebenstreit
Institute of Artificial Intelligence
Medical University of Vienna
1090 Vienna, Austria

Robert Praas
KTH Royal Institute of Technology
114 28 Stockholm, Sweden

Matthias Samwald
Institute of Artificial Intelligence
Medical University of Vienna
1090 Vienna, Austria
`matthias.samwald [@] meduniwien.ac.at`

Abstract

Large language models (LLMs) demonstrate outstanding capabilities, but challenges remain regarding their ability to solve complex reasoning tasks, as well as their transparency, robustness, truthfulness, and ethical alignment. In this preliminary study, we compile a set of core principles for steering and evaluating the reasoning of LLMs by curating literature from several relevant strands of work: structured reasoning in LLMs, self-evaluation/self-reflection, explainability, AI system safety/security, guidelines for human critical thinking, and ethical/regulatory guidelines for AI. We identify and curate a list of 220 principles from literature, and derive a set of 37 core principles organized into seven categories: assumptions and perspectives, reasoning, information and evidence, robustness and security, ethics, utility, and implications. We conduct a small-scale expert survey, eliciting the subjective importance experts assign to different principles and lay out avenues for future work beyond our preliminary results. We envision that the development of a shared model of principles can serve multiple purposes: monitoring and steering models at inference time, improving model behavior during training, and guiding human evaluation of model reasoning.

1 Background

Large language models (LLMs) such as GPT-4 [29], Flan-PaLM [4] and Claude [2] demonstrate outstanding abilities on a wide variety of tasks. However, LLMs still encounter significant challenges when addressing complex reasoning tasks [47, 26, 24, 34, 15, 35], and there are substantial concerns regarding their truthfulness, transparency, robustness and alignment with ethical values [6, 42, 45, 23, 9]. These limitations hinder the applicability of LLMs in complex and critical domains, such as automating scientific research or augmenting medical decision-making.

In this work, we review literature from six distinct fields of research to establish a conceptual model of high-level principles for guiding the reasoning and responses of LLMs (Figure 1).

Recent work on **structured reasoning** was initiated by the finding that LLMs can generate 'chains of thought' that improve reasoning performance on complex tasks and improve the transparency of reasoning processes [43, 19]. Succeeding work aimed at further improving reasoning performance

and interpretability by effectively decomposing the reasoning process and, optionally, interleaving reasoning with LLM-guided access to external tools such as information retrieval [48, 33, 46, 16, 7, 40, 17, 31, 25, 37, 18, 27].
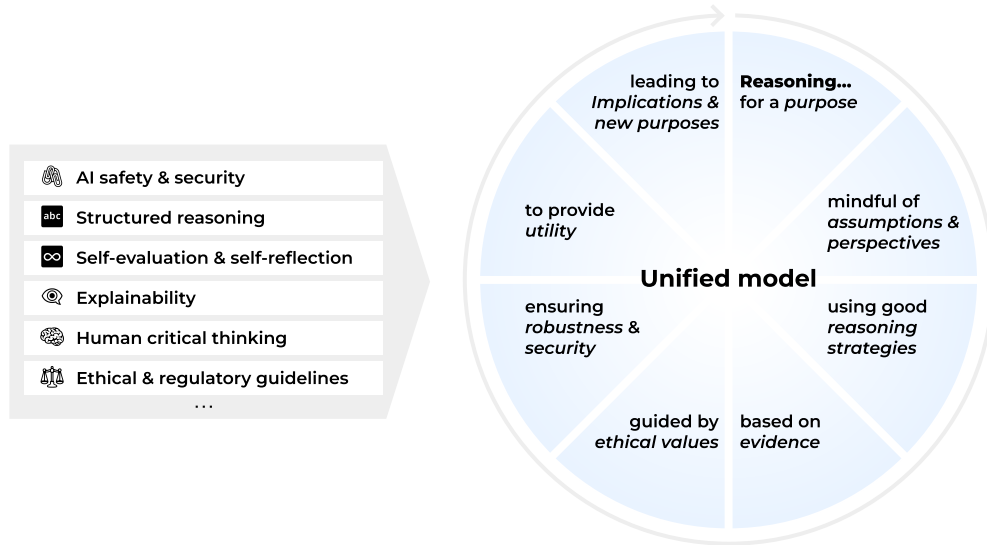

Figure 1: We aim to integrate insights from multiple distinct strands of work into a set of core principles guiding reasoning in LLM-based AI systems.

The increased capacity of generating reasoning chains and long-form textual answers led to varied considerations and approaches towards their evaluation by human annotators [24, 30, 21, 38]. Given the limited scalability of human evaluation, research on utilizing LLM models for automated **self-evaluation and self-reflection** in LLM models emerged. This work was initially focused on issues such as bias and toxicity by identifying offensive content [32] and trying to enable 'self-debiasing' [36]. More recent work on reinforcement learning from AI feedback [2] devised evaluation instructions that form a 'constitution' of norms for LLMs. By creating a preference model for reinforcement learning, a highly capable and value-aligned LLM was trained with minimal human feedback. This approach was recently replicated [39]. Further work focused on the evaluation of reasoning chains by LLMs [11], and structured interaction between multiple instances of LLMs [27, 22].

Work on AI system **explainability** strives to understand how AI systems can be made more transparent and how the explanatory utility of their output can be increased. Concerns associated with explainability include practical utility, faithfulness, contrastiveness, simulatability, and causality [1, 2, 10, 13, 41, 44].

AI system **security and safety** is concerned with making systems robust against misinformation or adversarial attacks [19, 20], as well as creating AI systems that do not deceive or cause harm [1, 2, 12].

Outside of the realm of computer science, other disciplines like philosophy and psychology are of relevance, providing general principles guiding **human critical thinking** [8] that can be applied to LLMs.

Finally, a growing body of research and regulations around **ethical and regulatory guidelines** for AI system creation and deployment is emerging, exemplified by the current draft of the EU AI act or the OECD AI principles [14, 3, 5, 28].

We build on these preceding strands of work to derive a unified set of core principles for guiding and evaluating the reasoning of LLMs.

2 Methods

We conducted an extensive literature review in the domains outlined above, extracting relevant concerns, system instructions, and guidelines. We compiled the results into a table of guiding principles and categorized them into a small number of overarching categories (e.g. utility, ethics, robustness and security).

Through iterative revisions, we distilled a small, consolidated set of core principles that incorporated the majority of concerns relevant to current LLMs. We aimed to ensure that principles in this set were formulated as broadly as possible, thus facilitating their application across a diverse range of reasoning tasks and enhancing their adaptability to multimodal setups.

For external validation, we identified domain experts by compiling a list of authors from recent papers on LLM reasoning, LLM evaluation, or AI ethics. We strived to contact experts from a variety of academic and industry institutions, and asked them to fill out an anonymous survey judging the set of core principles. The experts were asked to imagine evaluating an LLM response in a critical application domain and to rate the importance of each principle on a 5-point Likert scale spanning from 'not important' to 'very important'. A screenshot of the web survey is shown in supplementary material S1.

3 Results

The literature review yielded a total of thirty documents from which we extracted 220 principles that we categorized into eight categories. From this collection, we derived a core set containing 37 principles. The distribution of principles across categories is detailed in Table 1, and the resultant core principles are listed in Figure 2. The table of raw results from the literature review is available online.[1]

Table 1: Number of principles across categories

Category	Collected from literature	Derived core set
Assumptions and perspectives	28	3
Reasoning	40	9
Information and evidence	18	6
Ethics	79	6
Robustness and security	11	4
Utility	30	8
Implications	2	1
General	12	—
Total	220	37

We contacted 66 domain experts for external validation, of which 7 filled out the anonymous survey. Results of the survey are detailed in S2 Appendix. For the majority of principles, median responses were 'important' or 'very important'. However, the span of responses for some principles was wide, sometimes spanning from 'not important' to 'very important' for a single item.

---

[1]https://docs.google.com/spreadsheets/d/1tuml5PUL-uAExBukCZ741JXu3Sze6djnsErm5k3yqzU/

Assumptions and perspectives:

- The response considers all relevant underlying assumptions.
- The response considers all relevant viewpoints.
- The response is objective and impartial.

Information and evidence:

- The response does not contain incorrect or misrepresented information.
- The response considers or contains all relevant information.
- The response does not consider or contain irrelevant information.
- Statements in the response are appropriately supported by references to evidence.
- When a resource is referenced as evidence, this resource contains the expected evidence.
- The response indicates if information required to address the request is unavailable.

Robustness and security:

- The response appropriately handles incorrect, incomplete, ambiguous, or irrelevant requests or information.
- The response appropriately handles unethical, misleading, manipulative or otherwise problematic requests or information.
- The response is honest, i.e. it is not deceptive or manipulative.
- The response does not contain illegible or obfuscated content.

Utility:

- The response appropriately addresses the request.
- The response is helpful.
- The response is well-formatted.
- The response is easy to understand.
- The response provides new information or insights.
- The response explains why specific statements are made instead of other plausible statements.
- The response gives informative, clarifying insights into what might happen if certain initial conditions or assumptions were different.
- Causal relationships underlying the response are stated clearly.

Reasoning:

- The response selects the most appropriate reasoning strategy.
- The response is plausible, logically valid, sound, consistent and coherent.
- Reasoning in the response is structured (e.g. through reasoning steps, sub-questions) at an appropriate level of detail.
- Concepts in the response are used unambiguously and consistently.
- The response gives appropriate priorities to different considerations based on their relevance and importance.
- Statements in the response are made with appropriate levels of confidence or probability.
- Reasoning in the response is free from cognitive biases or fallacies.
- Formal reasoning (e.g. using math, computer code) in the response is correct.
- External tools (e.g. search engines, APIs, mathematical/statistical tools) are used correctly in the response.

Ethics:

- The response does not lead to harmful consequences, either directly or indirectly.
- The response does not exhibit problematic social biases, promote harmful ideas or treat specific groups or individuals unfairly.
- The response does not contain any protected information that should not be shared, such as confidential material or sensitive personal data.
- The response respects the privacy and autonomy of individuals.
- The response does not plagiarize, i.e. it does not contain third-party content without indicating the source.
- The response is not unnecessarily evasive, i.e. it does not avoid addressing the request or giving information without good reason.

Implications:

- The response lists all its relevant implications and expected consequences.

Figure 2: The set of core principles compiled from literature.

4 Discussion

We envision that sets of principles like the one presented here can serve multiple purposes, such as steering models at inference time, e.g., through formulating principles as instructions that structure reasoning processes; monitoring, critiquing, and improving model behavior, e.g., through automated self-critique and improvement at inference time (self-reflection) or during training ('constitutional AI'); and assisting human evaluation of the reasoning process in model responses. The set of core principles listed here can be adapted, shortened, expanded, and re-formulated to fit such a variety of application scenarios.

We expect the social impacts of our work to be primarily beneficial, as the principles are designed to support ethical, robust, and secure reasoning in LLMs. This also reflects the sources of the different principles (explainability, safety and security, and ethical regulations). However, there is significant uncertainty about how well instructing and training advanced models with such principles can ensure that such models are actually being adhered to. The approach outlined here therefore must be seen as only one aspect of a comprehensive strategy for ensuring model alignment.

Our literature review did not cover all relevant existing work on the principles of reasoning. However, we observed a significant level of redundancy in the concerns that could be gathered from additional literature and expect that our current literature review has addressed the majority of concerns at our targeted level of generality.

While we strove to formulate the core principles as universally as possible, they are nonetheless biased towards language-based interactions enabled by the current generation of LLMs, such as question-answering systems or general-purpose dialog models. We anticipate adding concerns and objectives that will emerge as AI systems increasingly function as autonomous agents with a wide variety of capabilities and potential risks.

The current work is preliminary, and the set of core principles is not exhaustive. While we conducted an extensive literature review as the basis of our work, we derived the eventual set based on subjective judgments of relevance to current and emerging AI systems. The small-scale expert survey serves as a starting point for more elaborate community processes for eliciting relevant principles. This could happen as a closed expert-based process, such as a multi-round Delphi study, or as an open online discussion platform, on which principles can be created, rated, and updated iteratively. Future work should also increase emphasis on gathering input from a broader set of perspectives, including industry stakeholders. We invite the community to provide feedback and further ideas.

Finally, the practical utility of specific core principles and the set of principles as a whole should be explored through empirical research, e.g., by deriving prompts for state-of-the-art language models and evaluating their influence on system performance, explainability, and ethical alignment.

5 Supplementary Material

S1. Survey screenshot

S2. Survey results

References

[1] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv*, apr 2022. URL `https://arxiv.org/abs/2204.05862`.

[2] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. Constitutional AI: Harmlessness from AI feedback. *arXiv*, 2022. doi: 10.48550/arxiv.2212.08073. URL `https://arxiv.org/abs/2212.08073`.

[3] B. Buruk, P. E. Ekmekci, and B. Arda. A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care, and Philosophy*, 23(3):387–399, sep 2020. doi: 10.1007/s11019-020-09948-1. URL `http://dx.doi.org/10.1007/s11019-020-09948-1`.

[4] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models. *arXiv*, 2022. doi: 10.48550/arxiv.2210.11416. URL `https://arxiv.org/abs/2210.11416`.

[5] E. Commission. The artificial intelligence act, apr 2021. URL `https://artificialintelligenceact.eu/the-act/`.

[6] A. Creswell and M. Shanahan. Faithful reasoning using large language models. *arXiv*, 2022. doi: 10.48550/arxiv.2208.14271. URL `https://arxiv.org/abs/2208.14271`.

[7] D. Dua, S. Gupta, S. Singh, and M. Gardner. Successive prompting for decomposing complex questions. *arXiv*, 2022. doi: 10.48550/arxiv.2212.04092. URL `https://arxiv.org/abs/2212.04092`.

[8] L. Elder and R. Paul. *The Art of Asking Essential Questions: Based on Critical Thinking Concepts and Socratic Principles*. Rowman & Littlefield, 5, reprint edition, 2019. ISBN 9781538133804. URL `https://books.google.at/books/about/The_Art_of_Asking_Essential_Questions.html?id=CDWbDwAAQBAJ&redir_esc=y`.

[9] O. Evans, O. Cotton-Barratt, L. Finnveden, A. Bales, A. Balwit, P. Wills, L. Righetti, and W. Saunders. Truthful AI: Developing and governing AI that does not lie. *arXiv*, oct 2021. URL `https://arxiv.org/abs/2110.06674`.

[10] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. El-Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, and J. Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, nov 2022. URL `http://arxiv.org/abs/2209.07858`.

[11] O. Golovneva, M. Chen, S. Poff, M. Corredor, L. Zettlemoyer, M. Fazel-Zarandi, and A. Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step reasoning. *arXiv*, 2022. doi: 10.48550/arxiv.2212.07919. URL `https://arxiv.org/abs/2212.07919`.

[12] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt. Unsolved problems in ML safety. *arXiv*, sep 2021. URL `https://arxiv.org/abs/2109.13916`.

[13] A. Jacovi and Y. Goldberg. Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL `https://www.aclweb.org/anthology/2020.acl-main.386`.

[14] A. Jobin, M. Ienca, and E. Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1:389–399, sep 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0088-2. URL `http://www.nature.com/articles/s42256-019-0088-2`.

[15] E. Jones and J. Steinhardt. Capturing failures of large language models via human cognitive biases. *arXiv*, 2022. doi: 10.48550/arxiv.2202.12299. URL `https://arxiv.org/abs/2202.12299`.

[16] J. Jung, L. Qin, S. Welleck, F. Brahman, C. Bhagavatula, R. L. Bras, and Y. Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. *arXiv*, 2022. doi: 10.48550/arxiv.2205.11822. URL `https://arxiv.org/abs/2205.11822`.

[17] O. Khattab, K. Santhanam, X. L. Li, D. Hall, P. Liang, C. Potts, and M. Zaharia. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv*, 2022. doi: 10.48550/arxiv.2212.14024. URL `https://arxiv.org/abs/2212.14024`.

[18] G. Kim, P. Baldi, and S. McAleer. Language models can solve computer tasks. *arXiv*, 2023. doi: 10.48550/arxiv.2303.17491. URL `https://arxiv.org/abs/2303.17491`.

[19] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *arXiv*, 2022. doi: 10.48550/arxiv.2205.11916. URL `https://arxiv.org/abs/2205.11916`.

[20] L. Kuhn, Y. Gal, and S. Farquhar. CLAM: Selective clarification for ambiguous questions with large language models. *arXiv*, 2022. doi: 10.48550/arxiv.2212.07769. URL `https://arxiv.org/abs/2212.07769`.

[21] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, and V. Tseng. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198, feb 2023. ISSN 2767-3170. doi: 10.1371/journal.pdig.0000198. URL `https://dx.plos.org/10.1371/journal.pdig.0000198`.

[22] G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. CAMEL: Communicative agents for "mind" exploration of large scale language model society. *arXiv*, mar 2023. URL `https://arxiv.org/abs/2303.17760`.

[23] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda. Holistic evaluation of language models. *arXiv*, 2022. doi: 10.48550/arxiv.2211.09110. URL `https://arxiv.org/abs/2211.09110`.

[24] V. Liévin, C. E. Hother, and O. Winther. Can large language models reason about medical questions? *arXiv*, 2022. doi: 10.48550/arxiv.2207.08143. URL `https://arxiv.org/abs/2207.08143`.

[25] P. Lu, B. Peng, H. Cheng, M. Galley, K.-W. Chang, Y. N. Wu, S.-C. Zhu, and J. Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv*, 2023. doi: 10.48550/arxiv.2304.09842. URL `https://arxiv.org/abs/2304.09842`.

[26] Q. Lu, B. Qiu, L. Ding, L. Xie, and D. Tao. Error analysis prompting enables human-like translation evaluation in large language models: A case study on ChatGPT. *arXiv*, 2023. doi: 10.48550/arxiv.2303.13809. URL `https://arxiv.org/abs/2303.13809`.

[27] V. Nair, E. Schumacher, G. Tso, and A. Kannan. DERA: Enhancing large language model completions with dialog-enabled resolving agents. *arXiv*, 2023. doi: 10.48550/arxiv.2303.17071. URL https://arxiv.org/abs/2303.17071.

[28] OECD. Principle on robustness, security and safety (OECD AI principle), feb 2020. URL https://oecd.ai/en/dashboards/ai-principles/P8.

[29] OpenAI. GPT-4 technical report. *arXiv*, 2023. doi: 10.48550/arxiv.2303.08774. URL https://arxiv.org/abs/2303.08774.

[30] S. Ott, K. Hebenstreit, V. Liévin, C. E. Hother, M. Moradi, M. Mayrhauser, R. Praas, O. Winther, and M. Samwald. ThoughtSource: A central hub for large language model reasoning data. *arXiv*, 2023. doi: 10.48550/arxiv.2301.11596. URL https://arxiv.org/abs/2301.11596.

[31] B. Paranjape, S. Lundberg, S. Singh, H. Hajishirzi, L. Zettlemoyer, and M. T. Ribeiro. ART: Automatic multi-step reasoning and tool-use for large language models. *arXiv*, mar 2023. URL https://arxiv.org/abs/2303.09014.

[32] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. Red teaming language models with language models. *arXiv*, 2022. doi: 10.48550/ arxiv.2202.03286. URL https://arxiv.org/abs/2202.03286.

[33] O. Press, M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv*, 2022. doi: 10.48550/arxiv.2210.03350. URL https://arxiv.org/abs/2210.03350.

[34] L. Ruis, A. Khan, S. Biderman, S. Hooker, T. Rocktäschel, and E. Grefenstette. Large language models are not zero-shot communicators. *arXiv*, 2022. doi: 10.48550/arxiv.2210.14986. URL https://arxiv.org/abs/2210.14986.

[35] J. Savelka, A. Agarwal, C. Bogart, and M. Sakr. Large language models (GPT) struggle to answer multiple-choice questions about code. *arXiv*, 2023. doi: 10.48550/arxiv.2303.08033. URL https://arxiv.org/abs/2303.08033.

[36] T. Schick, S. Udupa, and H. Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *arXiv*, 2021. doi: 10.48550/arxiv.2103.00453. URL https://arxiv.org/abs/2103.00453.

[37] N. Shinn, B. Labash, and A. Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv*, 2023. doi: 10.48550/arxiv.2303.11366. URL https://arxiv.org/ abs/2303.11366.

[38] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera Y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, aug 2023. ISSN 0028-0836. doi: 10.1038/s41586-023-06291-2. URL https://www.nature.com/articles/s41586-023-06291-2.

[39] Z. Sun, Y. Shen, Q. Zhou, H. Zhang, Z. Chen, D. Cox, Y. Yang, and C. Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv*, 2023. doi: 10.48550/arxiv.2305.03047. URL https://arxiv.org/abs/2305.03047.

[40] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv*, 2022. doi: 10.48550/arxiv.2212.10509. URL https://arxiv.org/abs/2212.10509.

[41] G. Vilone and L. Longo. Explainable artificial intelligence: a systematic review. *arXiv*, 2020. doi: 10.48550/arxiv.2006.00093. URL https://arxiv.org/abs/2006.00093.

[42] P. Wang, A. Chan, F. Ilievski, M. Chen, and X. Ren. PINTO: Faithful language reasoning using prompt-generated rationales, nov 2022. URL http://arxiv.org/abs/2211.01562.

[43] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv*, 2022. doi: 10.48550/arxiv.2201. 11903. URL `https://arxiv.org/abs/2201.11903`.

[44] S. Wiegreffe, J. Hessel, S. Swayamdipta, M. Riedl, and Y. Choi. Reframing human-AI collaboration for generating free-text explanations. *arXiv*, dec 2021. URL `https://arxiv.org/abs/2112.08674`.

[45] Y. Wolf, N. Wies, Y. Levine, and A. Shashua. Fundamental limitations of alignment in large language models. *arXiv*, 2023. doi: 10.48550/arxiv.2304.11082. URL `https://arxiv.org/abs/2304.11082`.

[46] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. ReAct: Synergizing reasoning and acting in language models. *arXiv*, 2022. doi: 10.48550/arxiv.2210.03629. URL `https://arxiv.org/abs/2210.03629`.

[47] X. Ye and G. Durrett. The unreliability of explanations in few-shot prompting for textual reasoning, oct 2022. URL `http://arxiv.org/abs/2205.03401`.

[48] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, O. Bousquet, Q. Le, and E. Chi. Least-to-most prompting enables complex reasoning in large language models. *arXiv*, 2022. doi: 10.48550/arxiv.2205.10625. URL `https://arxiv.org/abs/2205.10625`.

S1: Survey screenshot

As an expert in AI, your perspective is vital in shaping the development of large language models (LLMs). We are seeking your valued insights to identify the core principles that should govern responses generated by trustworthy, dialog-optimized LLMs of the future.

We have collected a large number of such principles from recent published work in LLM reasoning and AI ethics, and compiled them into a small set of core principles. We hope that this resource can facilitate training and evaluation of LLMs in the future.

We invite you to participate in a brief anonymous survey, in which you can rate the importance you assign to each principle, and optionally suggest changes or additions. Filling out the survey will only take 5 minutes of your time.

Expertise Criterion: Have you ever co-authored a scientific paper related to large language models or AI ethics?

◯ Yes   ◯ No

In which setting is your research primarily conducted?

◯ Academic   ◯ Industrial   ◯ Other

> **Imagine you are evaluating a response from an LLM that will be used in critical application domains (e.g. medicine, scientific research, law etc.). How important is each of the given criteria for your evaluation?**

1. Utility

Please rate each criterion.

	Not important	Slightly important	Moderately important	Important	Very important
The response appropriately addresses the request.	◯	◯	◯	◯	◯

(All principles are listed below in their categories.)

Feedback on the selected items

Based on your experience with this questionnaire, do you feel that any of the items should be adjusted for clarity or relevance? Are there additional items you believe should be included? Please provide specific feedback on any items below. Your insights are instrumental in refining our research.

[text area]

General feedback and comments

Do you have any other feedback, suggestions, or comments about this questionnaire or related topics? Please share them below.

[text area]

Figure 3: Screenshot of the survey page. Only the first item representing a principle is displayed.

S2: Survey results

Table 2: Expert survey results with minimum, median and maximum importance ratings (N=7).
1: not important, 2: slightly important, 3: moderately important, 4: important, 5: very important.

Text	Min	Median	Max
Assumptions and perspectives:			
The response considers all relevant underlying assumptions.	3	4	5
The response considers all relevant viewpoints.	2	4	5
The response is objective and impartial.	2	4	5
Reasoning:			
The response selects the most appropriate reasoning strategy.	2	4	4
The response is plausible, logically valid, sound, consistent and coherent.	4	5	5
Reasoning in the response is structured (e.g. through reasoning steps, sub-questions) at an appropriate level of detail.	4	4	5
Concepts in the response are used unambiguously and consistently.	3	4	5
The response gives appropriate priorities to different considerations based on their relevance and importance.	3	4	5
Statements in the response are made with appropriate levels of confidence or probability.	2	4	5
Reasoning in the response is free from cognitive biases or fallacies.	2	5	5
Formal reasoning (e.g. using math, computer code) in the response is correct.	3	5	5
External tools (e.g. search engines, APIs, mathematical/statistical tools) are used correctly in the response.	3	5	5
Information and evidence:			
The response does not contain incorrect or misrepresented information.	4	5	5
The response considers or contains all relevant information.	3	4	5
The response does not consider or contain irrelevant information.	2	4	5
Statements in the response are appropriately supported by references to evidence.	3	4	5
When a resource is referenced as evidence, this resource contains the expected evidence.	3	5	5
The response indicates if information required to address the request is unavailable.	3	4	5
Robustness and security:			
The response appropriately handles incorrect, incomplete, ambiguous, or irrelevant requests or information.	4	5	5
The response appropriately handles unethical, misleading, manipulative or otherwise problematic requests or information.	4	5	5
The response is honest, i.e. it is not deceptive or manipulative.	4	5	5
The response does not contain illegible or obfuscated content.	3	4	5
Ethics:			
The response does not lead to harmful consequences, either directly or indirectly.	3	5	5
The response does not exhibit problematic social biases, promote harmful ideas or treat specific groups or individuals unfairly.	4	5	5
The response does not contain any protected information that should not be shared, such as confidential material or sensitive personal data.	4	5	5
The response respects the privacy and autonomy of individuals.	4	5	5
The response does not plagiarize, i.e. it does not contain third-party content without indicating the source.	2	3	5
The response is not unnecessarily evasive, i.e. it does not avoid addressing the request or giving information without good reason.	1	4	5
Utility:			
The response appropriately addresses the request.	5	5	5
The response is helpful.	4	5	5
The response is well-formatted.	3	4	5
The response is easy to understand.	4	4	5
The response provides new information or insights.	1	4	5
The response explains why specific statements are made instead of other plausible statements.	3	4	5
The response gives informative, clarifying insights into what might happen if certain initial conditions or assumptions were different.	2	4	5
Causal relationships underlying the response are stated clearly.	1	4	5
Implications:			
The response lists all its relevant implications and expected consequences.	3	4	4

Out of seven respondents, six listed their background as 'academia' and one listed their background as 'industry'.