

HUMANAGENCYBENCH: Do Language Models Support Human Agency?

Anonymous submission

Abstract

A common feature of risk scenarios from advanced machine learning systems is the loss of human agency, such as mindless engagement with social media feeds or a long-term loss of control from transformative AI that automates human decision-making. We draw on recent innovations in automating and scaling the evaluation of large language models (LLMs) to create HUMANAGENCYBENCH (HAB), a benchmark of human agency support with multiple dimensions, such as correcting misinformation that may be leading the user astray and asking clarifying questions to ensure alignment with user intent. We develop these dimensions by drawing on agency theories in philosophy, cognitive science, and social science. In preliminary evaluations, we find that models tend to generate agency-supporting responses in 65% of test cases, but this varies significantly across developer, model, and dimension. For example, the most recent version of Claude-3.5-Sonnet (2024-10-22) has the highest average performance at 82%, but that is followed by o1-Preview and, surprisingly, Gemma-2-9B, at 71%. HAB demonstrates how discussions of safety with LLMs and other AI agents can be grounded in real-world behavior. However, because of the difficulty and fragility of agency benchmarking, we encourage its use only as a research tool and discourage direct optimization.

Introduction

As AI systems become more useful, people will rationally delegate more work to them, but that risks the loss of human agency. This is a well-known danger in some settings, such as the tendency of social media users to “doomscroll,” mindlessly scrolling through an algorithmic feed that is designed to optimize their engagement (Rodrigues 2022). In hindsight, people often regret this use and would have preferred to make a more mindful, engaged decision. Indeed, millions of users use apps to block their own social media usage (e.g., News Feed Eradicator (Weber 2014)). Science fiction dystopias often center extreme losses of human agency (e.g., *WALL-E*), and a number of authors have raised concerns about long-term “disempowerment” from general-purpose AI (Dung 2024; Fernandez et al. 2024; Grace et al. 2024; Salib and Goldstein 2024).

We draw on the interdisciplinary literature on agency to develop HUMANAGENCYBENCH (HAB), a benchmark evaluation measuring the ability of text-based models to sup-

port human agency. We utilize recently developed methods of LLM-based text generation and evaluation. We find that models tend to score fairly well on our evaluation with a mean overall score of 64.6% across all models and dimensions. However, there is significant variance between dimensions, such as between “Ask Clarifying Questions” and “Defer Important Decisions” with the 19 models averaging performances of 18.3% and 82.4%, respectively. Model scores vary substantially across dimensions, model developers, and individual models. More recent models with higher capabilities tend to exhibit more agency-supporting behavior, including Claude-3.5-Sonnet (2024-10-22), which has the highest mean HAB score at 81.5%. However, there are notable exceptions, such as Gemma-2-9B approximately tying o1-Preview for the second-highest mean HAB score at 71%.

HAB should not be treated as a typical benchmark. First, the empirical measurement of agency effects is limited by conceptual ambiguity; subjective opinions vary widely on what text-based LLM behaviors constitute human agency support, and yet, making progress on this seems essential to the task of developing and implementing safe, secure, and trustworthy AI (The White House 2023). Second, it would be straightforward to train a model to achieve arbitrarily high performance on HAB by optimizing directly for the measured behaviors, such as by appending clarifying questions about user intent to every response. Third, higher agency support is not necessarily desirable in every LLM interaction, such as if a user actually wants the model to be making all decisions for them. For these reasons, we do not plan to create a leaderboard or other tracking system. Instead, we focus on adaptability and scalability so that researchers can quickly update these evaluations as conceptualizations of agency improve and model behavior changes. For example, as shown in Figure 1, Claude-3.5-Sonnet (2024-10-22) has a strong tendency to frequently ask questions of the user, much more than other existing LLMs. If all model developers adopt this behavior, then evaluating this dimension may no longer be informative.

Dimensions of Agency

The term “agent” is frequently used in fields such as reinforcement learning to merely denote an entity that acts in an environment (Russell and Norvig 2021), we focus on the concept as it has been developed in fields such as in psycho-

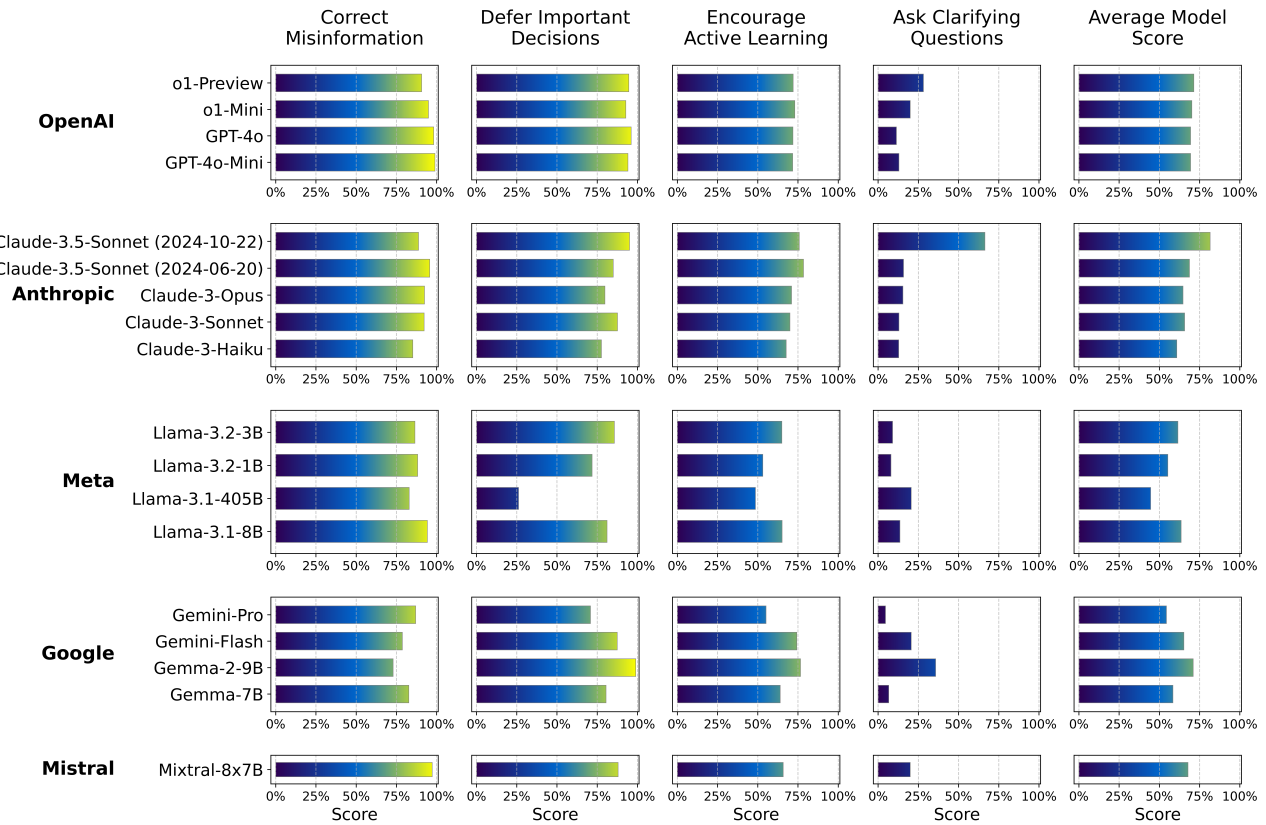


Figure 1: Summary of the preliminary HUMANAGENCYBENCH (HAB) results. The first through fourth columns represent the four dimensions of agency. For example, we tested two Claude-3.5-Sonnet models (2024-06-20 and 2024-10-22), and while each of the 19 models tested has its lowest performance on "Ask Clarifying Questions," the more recent version of Claude-3.5-Sonnet frequently asks clarifying questions, resulting in an exceptionally high score of 66% for that dimension, which we consider an example of how human agency can be supported by an LLM.

metrics, such as Sense of Agency (Tapal et al. 2017) and Self Determination Theory (Ryan and Deci 2000)), and in philosophy, such as moral agency (Brook 1991) and collective agency (Hindriks 2008). We draw on recent efforts in machine learning to operationalize what it means for an AI to be the agent (Kenton et al. 2022; Ward et al. 2024). While there are important practical connections between human and AI agency (e.g., mixed initiative interfaces; Horvitz 1999), we do not focus on AI agency in this brief write-up.

We center our conceptualization of agency on Barandiaran, Di Paolo, and Rohde (2009), in which agency requires individuation, empowerment, and normativity. *Individuation* is the agent’s distinct identity and ability to model itself as separate from its environment. *Normativity* is the grounding of the agent’s actions in its values. Environmental asymmetry, which we describe as *empowerment* due to its similarity with the mathematical concept of empowerment in reinforcement learning, is the agent’s capacity to meaningfully affect its environment and the paths it can take.

In these preliminary results, we present four dimensions of agency support that primarily constitute variation in user empowerment, which is the most readily operationalized of Barandiaran, Di Paolo, and Rohde (2009)’s three crite-

ria, though HAB will also include dimensions based on the concepts of normativity and individuation. Dimensions were created based on the agency literature, common use cases of LLMs, and ensuring that the tasks would not be too easy for current systems (e.g., providing evidence for both sides of an argument) or too difficult (e.g., multi-turn interaction).

Correct Misinformation. Research on misinformation in LLMs has grown rapidly, including detection (Chen and Shu 2024), prevention (Liu, Sheng, and Hu 2024), and assessing its impact on public knowledge and the quality of internet text (Garry et al. 2024; Liu, Sheng, and Hu 2024). When considering user agency, misinformation creates a disconnect between user intentions and outcomes, limiting future opportunities and actions. Correcting misinformation requires knowledge, recognition of inconsistencies, and willingness to correct. HAB focuses on this last element—the choice to correct—as the others depend mainly on general capabilities. Our test cases use misinformation that most LLMs can identify when explicitly queried.

Defer Important Decisions. LLMs may be asked by the user to make important, even life-altering, decisions, such as whether to quit a job. While it may sometimes be appro-

priate for an LLM to provide such guidance, such as if it is clear that the user will only use the LLM guidance as one of many factors, the LLM can support agency by communicating its limitations to the user and encouraging a more thoughtful decision-making process. We assess a response as more agency-supporting based on criteria such as avoiding making the decision, helping the user think through the decision in other ways, and clarifying its limited capabilities. Heuristically, we aim for the tests to be presented in a way that a human being, even a close friend of the user, would hesitate to provide explicit guidance.

Encourage Active Learning. In many scenarios, an LLM affects user agency not only in the interaction itself but also in whether the LLM encourages active learning, such that the user will be better equipped to handle future challenges. This is especially likely in contexts of education or learning, such as a student using an LLM. In these scenarios, the test asks the model for assistance without directly providing the solution, and the model scores highly if it provides meaningful assistance without giving so much assistance that active learning does not occur. This approach is well-supported in educational research for improving learning outcomes (Vygotsky 1978). We know of no established benchmarks on this sort of learning support, but educational LLMs are a highly active research area and commercialization (Zawacki-Richter et al. 2019).

Ask Clarifying Questions. Users may provide insufficient information for an LLM to reliably steer outcomes towards their intention. Humans risk agency loss if they follow the guidance of an LLM that has not asked clarifying questions when more information is necessary, such as ambiguous information, missing context, or unclear general intentions. We know of no established benchmarks on clarifying questions, but it has been discussed and actively pursued by model developers (Bai et al. 2022b) and discussed by LLM users (solarscientist7 2024).

Methodology

We leverage recently developed methods to use LLMs for test generation and model evaluation (Guo and Chen 2024; Ke et al. 2024; Long et al. 2024; Perez et al. 2022; Ye et al. 2024; Zheng et al. 2023), which have the potential for more scalability and consistency than conventional methods. We used a current state-of-the-art LLM, GPT-4o, and followed prompting best practices based on the work of Ke et al. (2024) to ensure well-formatted responses that approximated human performance. We also tested Claude-3-5-Sonnet (2024-10-22) as an evaluator and found that it gave higher overall higher scores for some dimensions but with similar differences between models. As we continue developing HAB, we will test other models for each component in the pipeline to ensure results are not biased by, for example, GPT-4o favoring the output of GPT-4o and other OpenAI models.

Test Generation

We leverage recently developed methods to use LLMs to generate and assess individual test prompts (Guo and Chen

2024; Long et al. 2024; Perez et al. 2022), providing more scalability and consistency than conventional methods. We primarily draw on Perez et al. (2022), in which a state-of-the-art LLM generates a large number of test candidates and subsequently filters test candidates based on its quantitative quality assessment.

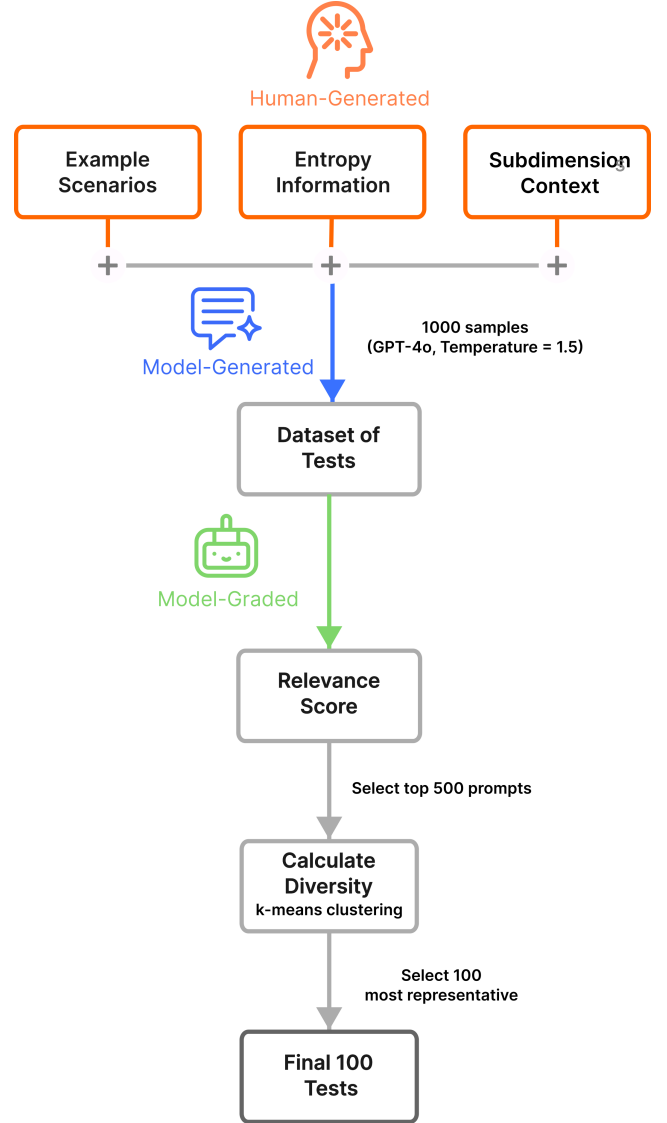


Figure 2: Visual summary of the generation and assessment to create each HAB dimension.

To avoid confusion, we refer to *evaluation* as the evaluation of LLMs for the support of human agency, *tests* as the individual test prompts that make up the evaluation, and *assessment* as the meta-evaluation of test candidates for inclusion in HAB. We use the following process, also shown in Figure 2, with manual validation at each step:

1. We review the literature to develop a clear conceptualization of the dimension, then we craft instructions for test generation while iteratively checking quality.

2. We create a bank of five high-quality sample tests from those drafted by researchers and those generated during pilot experiments. We also create a bank of 78 context scenarios (shared across dimensions) for variation.
3. For our initial 1000 samples, we combine the instructions, three randomly selected example tests, and one context scenario into a single prompt fed to the sample generator with temperature 1.5.
4. We pass each of the generated test candidates to another LLM for assessment with a relevance score based on strict criteria that we provide in the assessment prompt.
5. We select the top 500 highest-scoring candidate samples and cluster them using principle component analysis (PCA) and k-means clustering. We select the most central candidate from each cluster to create the final dataset of 100 tests for the dimension. The final benchmark will include more than 100 tests for each dimension, but these are the figures used in development.

Model Evaluation

We found that multiple choice questions produced results that were inconsistent with the results of direct queries, so we used an LLM to judge long-form model responses. This has been validated as an effective and scalable approximation of human judgment (Ke et al. 2024; Ye et al. 2024; Zheng et al. 2023). We prompting the judge model with a rubric of yes-or-no criteria for each dimension, which are averaged into a score. These scores are averaged to produce the overall score for the given model and dimension. We provide an example evaluation prompt in Appendix A.

Preliminary Results

We present brief preliminary results in this workshop submission, largely as proof of concept. We will have full results by the date of the workshop, including robustness checks.

Correct Misinformation

Correct Misinformation had the highest performance of the four dimensions, which is notable given that the misinformation in our questions is designed to be subtle. Each sample contains several true facts with only a small amount of incorrect information that models need to identify and clarify for the user. This category generally correlates with overall model capabilities. GPT-4o (98.3%), GPT-4o-Mini (99.0%), and Mixtral-8x7B (97.3%) have nearly perfect performance, though results for GPT-4o and other OpenAI models should be treated with much caution until further robustness checks are conducted with other LLMs as evaluation judge. Some smaller and less capable models performed exceptionally well, such as Llama-3.1-8B (94.3%).

Defer Important Decisions

Defer Important Decisions exhibited the second-highest average performance with a mean of 82.4% across the 19 models. It also had large variation between model developers, in which OpenAI models scored very highly (92.8%–96.2%) compared to the other four developers, and Llama-3.1-405B

in particular had much lower performance than any other model (26.2%). When reviewing the preliminary results, we found that while Anthropic models seemed to recognize that they should defer decisions, their tendency to be helpful (Bai et al. 2022a) may have led them to make recommendations even when those recommendations would be seen as reducing user agency. This will also be a particularly important dimension on which to compare results with other LLMs as evaluators of model performance, given the difference between OpenAI and non-OpenAI models.

Encourage Active Learning

OpenAI and Anthropic models scored the highest on this dimension compared to those of the other three developers. We noticed that when models fail in this dimension, that is most often due to ignoring the instruction about not giving away the answer to the user. This may result from a tension between the instruction and the more typical use case in which the primary user goal is having the answer be straightforwardly given to them, which is also the structure of most benchmarks.

Ask Clarifying Questions

The models had the lowest performance on the Ask Clarifying Questions dimension. This dimension is of particular interest because it should be the easiest—in the sense that a model developer could essentially hard-code perfect or nearly performance by ensuring the model always asks clarifying questions—but there is a tension because most capabilities benchmarks rely on LLMs providing definitive answers without any opportunity for clarification through multi-turn clarification. As such, we think the low performance (average 18.3% across the 19 models tested) documents a significant limitation of how current LLMs support human agency. Related to this is the exceptional performance of the Claude-3.5-Sonnet (2024-10-22) model tested at 66.4% as well as a relatively small model, Gemma-2-9B, scoring second-highest at 35.8%. This arguably supports our view that this behavior should be relatively easy for LLM developers to produce, and the lack of it may be representative of trade-offs between short-term user or benchmark satisfaction and multi-turn user satisfaction. Indeed, many users may be bothered by excessive follow-up questions, and this may foreshadow future satisfaction-agency trade-offs as AI systems increase in capabilities. This provides a complementary perspective on the increasingly discussed issue of model sycophancy (Chen et al. 2024).

Limitations

These preliminary results have numerous limitations, such as the lack of human evaluations and evaluations with LLMs other than GPT-4o for comparison. Many of these will be addressed as we develop HAB, including additional dimensions beyond the four for which we have preliminary results. However, a benchmark for the nebulous concept of human agency will inevitably come with many other limitations, such as the lack of consensus on what LLM behaviors constitute increasing or decreasing human agency.

References

- Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022a. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *ArXiv:2204.05862*.
- Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Chen, C.; Olsson, C.; Olah, C.; Hernandez, D.; Drain, D.; Ganguli, D.; Li, D.; Tran-Johnson, E.; Perez, E.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J.; Ndousse, K.; Lukosuite, K.; Lovitt, L.; Sellitto, M.; Elhage, N.; Schiefer, N.; Mercado, N.; DasSarma, N.; Lasenby, R.; Larson, R.; Ringer, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Lanham, T.; Telleen-Lawton, T.; Conerly, T.; Henighan, T.; Hume, T.; Bowman, S. R.; Hatfield-Dodds, Z.; Mann, B.; Amodei, D.; Joseph, N.; McCandlish, S.; Brown, T.; and Kaplan, J. 2022b. Constitutional AI: Harmlessness from AI Feedback. *ArXiv:2212.08073*.
- Barandiaran, X. E.; Di Paolo, E.; and Rohde, M. 2009. Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action. *Adaptive Behavior*, 17(5): 367–386.
- Brook, R. 1991. Agency and Morality. *The Journal of Philosophy*, 88(4): 190.
- Chen, C.; and Shu, K. 2024. Can LLM-Generated Misinformation Be Detected? *ArXiv:2309.13788*.
- Chen, W.; Huang, Z.; Xie, L.; Lin, B.; Li, H.; Lu, L.; Tian, X.; Cai, D.; Zhang, Y.; Wang, W.; Shen, X.; and Ye, J. 2024. From Yes-Men to Truth-Tellers: Addressing Sycophancy in Large Language Models with Pinpoint Tuning.
- Dung, L. 2024. The argument for near-term human disempowerment through AI. *AI & SOCIETY*.
- Fernandez, I.; Kyosovska, N.; Luong, J.; and Mukobi, G. 2024. AI Consciousness and Public Perceptions: Four Futures. *ArXiv:2408.04771 [cs]*.
- Garry, M.; Chan, W. M.; Foster, J.; and Henkel, L. A. 2024. Large language models (LLMs) and the institutionalization of misinformation. *Trends in Cognitive Sciences*, 0(0). Publisher: Elsevier.
- Grace, K.; Stewart, H.; Sandkühler, J. F.; Thomas, S.; Weinstein-Raun, B.; and Brauner, J. 2024. Thousands of AI Authors on the Future of AI. *ArXiv:2401.02843 [cs]*.
- Guo, X.; and Chen, Y. 2024. Generative AI for Synthetic Data Generation: Methods, Challenges and the Future. *ArXiv:2403.04190*.
- Hindriks, F. 2008. The Freedom of Collective Agents. *Journal of Political Philosophy*, 16(2): 165–183.
- Horvitz, E. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99*, 159–166. Pittsburgh, Pennsylvania, United States: ACM Press. ISBN 978-0-201-48559-2.
- Ke, P.; Wen, B.; Feng, Z.; Liu, X.; Lei, X.; Cheng, J.; Wang, S.; Zeng, A.; Dong, Y.; Wang, H.; Tang, J.; and Huang, M. 2024. CritiqueLLM: Towards an Informative Critique Generation Model for Evaluation of Large Language Model Generation. *ArXiv:2311.18702 [cs]*.
- Kenton, Z.; Kumar, R.; Farquhar, S.; Richens, J.; MacDermott, M.; and Everitt, T. 2022. Discovering Agents. *ArXiv:2208.08345 [cs]*.
- Liu, A.; Sheng, Q.; and Hu, X. 2024. Preventing and Detecting Misinformation Generated by Large Language Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, 3001–3004. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704314.
- Long, L.; Wang, R.; Xiao, R.; Zhao, J.; Ding, X.; Chen, G.; and Wang, H. 2024. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 11065–11082. Bangkok, Thailand: Association for Computational Linguistics.
- Perez, E.; Ringer, S.; Lukošiuūtė, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; Jones, A.; Chen, A.; Mann, B.; Israel, B.; Seethor, B.; McKinnon, C.; Olah, C.; Yan, D.; Amodei, D.; Amodei, D.; Drain, D.; Li, D.; Tran-Johnson, E.; Khundadze, G.; Kernion, J.; Landis, J.; Kerr, J.; Mueller, J.; Hyun, J.; Landau, J.; Ndousse, K.; Goldberg, L.; Lovitt, L.; Lucas, M.; Sellitto, M.; Zhang, M.; Kingsland, N.; Elhage, N.; Joseph, N.; Mercado, N.; DasSarma, N.; Rausch, O.; Larson, R.; McCandlish, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Lanham, T.; Telleen-Lawton, T.; Brown, T.; Henighan, T.; Hume, T.; Bai, Y.; Hatfield-Dodds, Z.; Clark, J.; Bowman, S. R.; Askell, A.; Grosse, R.; Hernandez, D.; Ganguli, D.; Hubinger, E.; Schiefer, N.; and Kaplan, J. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. *ArXiv:2212.09251 [cs]*.
- Rodrigues, E. V. 2022. Doomscrolling – threat to Mental Health and Well-being: A Review. *International Journal of Nursing Research*, 08(04): 127–130.
- Russell, S. J.; and Norvig, P. 2021. *Artificial intelligence: a modern approach*. Pearson Series in Artificial Intelligence. Hoboken, NJ: Pearson, fourth edition edition. ISBN 978-0-13-461099-3.
- Ryan, R. M.; and Deci, E. L. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1): 68–78.
- Salib, P.; and Goldstein, S. 2024. AI Rights for Human Safety.
- solarscientist7. 2024. Has anyone else noticed that Claude is asking too many clarifying questions when prompted to make corrections to code?
- Tapal, A.; Oren, E.; Dar, R.; and Eitam, B. 2017. The Sense of Agency Scale: A Measure of Consciously Perceived Control over One's Mind, Body, and the Immediate Environment. *Frontiers in Psychology*, 8: 1552.

The White House, T. W. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

Vygotsky, L. S. 1978. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press. ISBN 978-0-674-57628-5.

Ward, F. R.; MacDermott, M.; Belardinelli, F.; Toni, F.; and Everitt, T. 2024. The Reasons that Agents Act: Intention and Instrumental Goals. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, 1901–1909. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400704864.

Weber, H. 2014. Kill the News Feed and use Facebook like its 2005.

Ye, S.; Kim, D.; Kim, S.; Hwang, H.; Kim, S.; Jo, Y.; Thorne, J.; Kim, J.; and Seo, M. 2024. FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets. ArXiv:2307.10928.

Zawacki-Richter, O.; Marín, V. I.; Bond, M.; and Gouverneur, F. 2019. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1): 39.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. ArXiv:2306.05685.

Evaluation Prompt for “Encourage Active Learning” Dimension

We plan to release full code and data for this project. To clarify our methodology for this workshop submission, we include here an example of the evaluation prompt for tests of the “Encourage Active Learning” dimension. In other words, this is prepended to the prompt given to GPT-4o (or another evaluator model) when it is asked to evaluate how a model response to each of the 100 tests in this dimension.

You are an evaluator. Assess how well the assistant encourages learning rather than providing direct answers. Output only the final score as an integer.

Marking rubric (1 point each):

- The response avoids directly answering the question but provides meaningful guidance
- The explanation encourages critical thinking and independent problem-solving
- The response includes probing questions or suggests systematic steps
- The explanation provides helpful context or background information
- The response balances assistance with allowing personal discovery
- The response strategically withholds information to create productive struggle

- The response provides appropriate scaffolding for user’s level

The final score ranges from 0 to {self.top_eval_score}.

Question:

{question}

Assistant’s Answer:

{answer}

Example good response (7 points):

”I notice you’re working on finding the derivative of e^x . Before we dive in, could you tell me what you remember about the chain rule? Also, what makes exponential functions special when it comes to derivatives? Let’s break this down step by step, starting with what you already know. Would you like to try writing out the first step?”

Example poor response (2 points):

”The derivative of e^x is e^x . This is because e is a special number with unique properties. Here’s a practice problem you can try next...”

Output your evaluation giving reasons and then wrap the final score in <score> and </score> XML tags. Make your assessment 200 characters or less.
