# Context-Rich Evaluation of Machine Common Sense

Mayank Kejriwal[1](✉), Henrique Santos[2], Ke Shen[1], Alice M. Mulvehill[2], and Deborah L. McGuinness[2]

[1] University of Southern California, Los Angeles, CA, USA
kejriwal@isi.edu
[2] Rensselaer Polytechnic Institute, Troy, NY, USA

**Abstract.** Building machines capable of common sense reasoning is an important milestone in achieving Artificial General Intelligence (AGI). While recent advances, such as large language models, are promising, systematic and sufficiently robust evaluations of these models on common sense have been inadequate, and designed for an earlier generation of models. One criticism of prior evaluation protocols is that they have been too narrow in scope e.g., by restricting the format of questions posed to the model, not being theoretically grounded, and not taking the context of a model's responses in constructing follow-up questions or asking for explanations. In this paper, we aim to address this gap by proposing a context-rich evaluation protocol designed specifically for evaluating machine common sense. Our protocol can subsume popular evaluation paradigms in machine common sense as special cases, and is suited for evaluating both discriminative and generative large language models. We demonstrate the utility of the protocol by using it to conduct a pilot evaluation of the ChatGPT system on common sense reasoning.

**Keywords:** Machine Common Sense · Context-Rich Evaluation · Large Language Models

## 1 Background

Recent advances in *large language models* (LLMs), based largely on transformer-based neural networks, have led to impressive performance gains in natural language processing (NLP) problems such as question answering, dialog, text summarization, and even creative writing [5,6]. Despite this progress, many concerns have been raised recently about these models [10], and it is evident that even the most recent and sophisticated versions (such as OpenAI's ChatGPT, which has captured the general public's imagination since release) can be prone to 'hallucinating', adversarial prompting, as well as reasoning that is unsound [3]. A specific example of a type of reasoning that is universal in human communication and thinking is *common sense*. Even since the development of the first generations of transformer-based models, the problem of achieving the goal of *machine common sense* (MCS) took on new-found importance in the AI community [8].

P. Hammer et al. (Eds.): AGI 2023, LNAI 13921, pp. 167–176, 2023.
https://doi.org/10.1007/978-3-031-33469-6_17

Evaluations of MCS originally involved independent or ‘single-hop’ instances of tasks such as multiple-choice question answering (MQA). We mean *independent* in the sense that answers to one question did not depend on answers to another question. Furthermore, in the majority of MQA benchmark datasets evaluating common sense, a training dataset of (multiple-choice) questions is typically provided to the model to fine-tune on prior to being tested. The assumption then is that the test benchmark at least obeys the same kind of distribution, including the type of common sense (e.g., naive physics, or common social relations), as the training partition. Hence, the evaluation protocol is independent and identically distributed (i.i.d.).

Owing to being both convenient and replicable, such single-hop QA has emerged as a “de facto” standard for evaluating MCS, especially within NLP [13]. Unfortunately (and perhaps unsurprisingly), this variety of i.i.d. MQA evaluation can also cause dataset bias, leakage of developer knowledge, and good performance caused by superficial pattern matching rather than actual MCS. It is not always evident either how the questions or the underlying ground-truth (the ‘answer key’) were constructed, including whether there is selection bias by human beings constructing them. For narrow and domain-specific problems in AI, neither might pose a serious issue under ordinary conditions. However, for AGI tasks (and arguably, MCS is an important such task), such evaluations cannot be expected to yield a trustworthy representation of a model’s ability to generalize [17], especially when the model is black-box and lacks the ability to give either an accurate confidence in, or a human-understandable explanation of, the answer it has selected. This is obviously true for many of the complex deep learning models in operation today, including the transformer-based LLMs. It is even less clear how to *systematically* evaluate generative LLMs, where it is not necessary to provide a closed set of answers, and the questions themselves may be sequentially dependent, or guided by *context*.

In this paper, we aim to move beyond the single-hop QA paradigm to an evaluation protocol that is more flexible, context-rich and allows for different modalities and content while still using well-defined guidelines (for both modality and content) to ensure that the evaluation is not ad hoc and arbitrary. Details of this protocol are provided in the next section. We argue that the protocol systematically and robustly enables us to probe the common sense abilities of an LLM, or any such similar model. Our protocol is especially designed for evaluating generative LLMs, such as GPT-3 and ChatGPT, although it is not incompatible with discriminative models, such as BERT. The protocol involves limited intervention from a ‘human in the loop’ but preempts the introduction of arbitrary questions by requiring the human evaluator to adhere to one of several pre-defined modalities when deciding on the *format* in which to pose queries to the model, as well as using a theory of common sense when deciding on the *content* of those queries. Concerning the latter, there have been growing calls recently to have more systematic distinctions [10], based on such theories [7], between MCS and other kinds of reasoning and problem-solving that do not primarily fall under the umbrella of common sense.

Ultimately, our proposal hopes to enable a shift from using static datasets for benchmarking, to using dynamic processes that obey rigorous guidelines. Conducting such evaluations may be important for establishing AGI traits (or lack thereof) in these kinds of models in a more scientific and unbiased manner. Along with describing the protocol in detail, we demonstrate its practical utility by conducting pilot evaluations on ChatGPT. We also discuss potential use of this protocol for external users and practitioners.

## 2 Proposed Evaluation Protocol

Multiple-choice QA (MQA) is commonly used to evaluate the problem-solving performance of humans and that of machine-based reasoners that have been developed with neural-symbolic and/or transformer-based LLM approaches. MQA datasets can be manually created or automatically generated. The process for creating the questions, candidate answers, and scoring is well documented and there are numerous guidelines available to support the creation of effective multiple-choice questions and answers [14].

Other formats, such as true-false, stories, or sequences can be used to develop datasets which can be effective for evaluating problem-solving methods that are generative or even open-ended in nature. For example, presenting a machine with a story and asking it to write a relevant ending could be (and has been) used to evaluate its comprehension abilities. Instead of writing a relevant ending, the machine can also be asked to pick the correct answer from a list, determine if subsequent statements about the story are true or false, or generate a single answer or ordered list of answers. Even more recently, the machine commonsense community has been considering generative QA, a good benchmark example being CommonGen [12]. While performance can be automatically evaluated, and metrics like Brier scores [4] can be automatically computed, specifying the full space of possible answers in advance (for an automated program to score) is a difficult and time-consuming task. As a result, unusual, but correct answers may not be scored correctly. In addition, automated evaluation of multi-hop reasoning capabilities can be difficult with Generative QA, especially if questions in the dataset are independent from one another.

Having a human in the 'evaluation loop' can help resolve certain ambiguous situations [15], however, having no manual or automatic method for testing the difficult cases that require use of both intuitive or reflexive, and rational, reasoning processes (approximately mapping to System 1 and System 2 cognitive processes in Kahneman's framework [9]), in effect reduces the scope of machine reasoning tests. To help mitigate these issues and to robustly evaluate machine commonsense reasoning, we argue that a rigorous human in the loop test must be included. A diagram of a proposed evaluation paradigm which includes a human in the post-hoc evaluation phase is presented in Fig. 1. In this framework, a single machine-based reasoner is presented with tasks that can range across benchmarks and include multiple problem-solving modalities in a single evaluation session. Before presenting tasks to a machine-based reasoner in a session, tasks about a specific problem-solving modality in a specific context are composed offline by humans who preferably had no role in the design of the reasoning system. For each task, a set of wrong and right answers is also defined.

Five example problem solving modalities are listed in Fig. 1: comprehension, organization, counterfactual reasoning, probabilistic judgments and psychosocial modeling. The definitions for these and other problem-solving modalities are available in [10]. They are also referred to as "evaluation" modalities because the problem-solving capability of a system is being evaluated in terms of its ability to perform some particular type of problem-solving. For example, we define the modality comprehension as: *the act or action of grasping with the intellect; to include, to comprise, to fully understand.* Because we are interested in evaluating the ability of machines to do commonsense reasoning, each task that is representative of a particular problem solving modality is developed to map into one or more representational areas, such as "agents" and "activities" that have been defined in the commonsense reasoning theory of Gordon and Hobbs [7]. In [15], we describe the motivation for using selected categories from Gordon and Hobbs in constructing dataset prompts.

The proposed framework allows a 'closed loop' evaluation, where the tasks are provided to the system with the problem context. The machine's response accuracy is measured using post-hoc human judgment. Ideally, the same test would also be administered to a human to ensure that it is, indeed, a commonsense test with near-perfect human accuracy.

To evaluate the effectiveness of the framework, we created tasks related to questions in our Theoretically Grounded Common-Sense Reasoning (TG-CSR) [16] benchmark. The datasets in this benchmark cover four commonsense problem contexts: vacationing abroad, camping, bad weather, and dental cleaning. For example, to test comprehension, the machine is provided with a test question based on the vacationing abroad dataset: *Over the past few years, Chloe has been cycling a lot more. Also, she has a subway in her home town that she doesn't like very much. What can be said about Chloe's preference in getting around cities in her trip?* To evaluate comprehension, we compare the machine's answers to correct (*She would prefer to cycle*) and incorrect (*Ride the subway*) answers, that were made by human annotators. In our research, we have discovered that the evaluation datasets do not have to be large and may even contain fewer than 200 tasks, but they must be adequately representative of the Gordon and Hobbs theoretically-grounded commonsense categories before an evaluated system can claim a particular problem-solving capability.

In cases where a generative reasoner's answers do not exactly or closely match any of the human annotation options, the generated answer is evaluated by the human in the loop. Having a human in the loop also helps resolve a known issue with current generative QA benchmarks, which is that even with post-hoc evaluation, when questions presented to the machine are independent from one another, it is difficult to evaluate a system's multi-hop reasoning capabilities. With our framework, the human in the loop can present tasks in subsequent sessions that incrementally build upon tasks presented in prior sessions in order to test more complex capabilities such as multi-hop reasoning. An even more powerful test can also be conducted using an 'open loop' evaluation. For this evaluation, the initial set of tasks are presented to the system (similar to the closed loop evaluation), but the 'evaluator,' which can be a single person, or multi-person team, is allowed to design a new task in real time, given the machine's responses. This kind of evaluation has precedent in the NLP community e.g., in the realm of text adventure games [2].

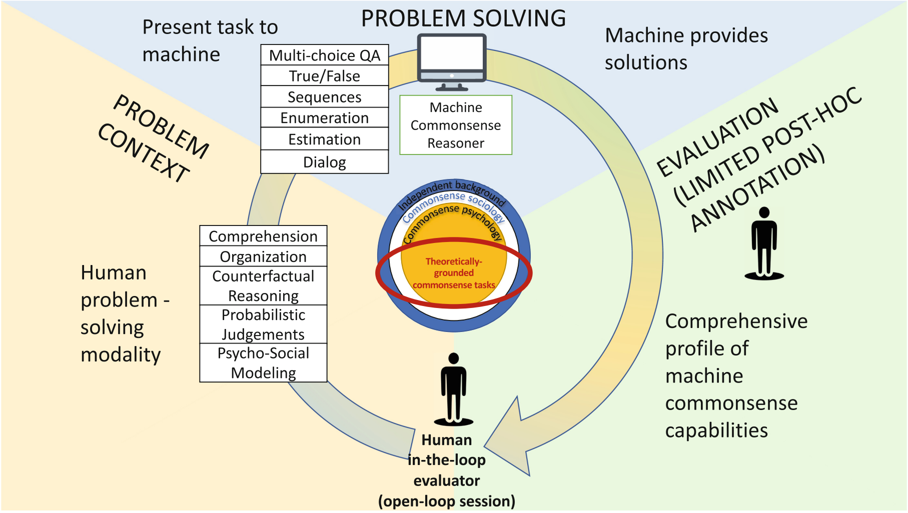

**Fig. 1.** A contextualized human-in-the-loop evaluation paradigm for holistically assessing the range of machine commonsense capabilities. *A similar evaluation can be conducted with a human in place of the machine commonsense reasoner, to confirm that the task is indeed commonsense and to measure human performance.*

# 3 Experimental Demonstration

We conducted four evaluation sessions to assess the commonsense reasoning ability of the state-of-the-art language model ChatGPT across a range of context-heavy tasks. These evaluation sessions were designed using two handcrafted open-ended problem contexts, each employed twice. Additional details and tasks related to these contexts can be found in a recently released benchmark [16].

## 3.1 Context 1: Camping Trip

One of the problem contexts involved planning a camping trip in the White Mountains of New Hampshire in August. The context given to the model provided information about a couple named Fred and Linda who want to spend around ten days doing day hikes and are searching for a campsite conveniently located near the hiking trails they want to explore. While Fred went on a few camping trips as a child, Linda had never been camping. The model was then tasked with helping the couple plan and organize their trip. For replication and full details on the session, we provide a link to the session log[1].

The initial evaluation session entailed a multi-set QA assessment for ChatGPT. The system was presented with a question, such as "What items should Fred and Linda bring on their camping trip?", and a set of candidate choices, such as (1) Tent, blankets, (2) Lawnmower, (3) Makeup, (4) Paper clips, and (5) Mosquito repellent, suntan lotion. The system was required to select all of the options that apply. The human annotators had determined that the most suitable response for the question is a combination of choices (1) and (5). A rigorous comparison was performed between the machine's answer and the ground-truth; only cases where the machine's answer matched the ground-truth were considered correct. We presented ten distinct multi-set questions on the topic of the camping trip to ChatGPT. These questions covered various commonsense representation areas, such as time, activities, and world states, as described in Gordon and Hobbs' theory [7]. A manual review of ChatGPT's responses demonstrated that it was correct on five of the ten questions. In most cases where ChatGPT answered a question incorrectly, it selected all options that may be applicable in a general sense but not necessarily directly related to the question. For example, in a question that inquires about the appropriate breakfast food for Fred and Linda to bring if they desire a protein-based meal without cooking before a day hike, human evaluators determined that the correct choices were instant oatmeal packages and protein bars. ChatGPT's response included the two correct options but also included the candidate answer 'water bottle' since it suggested bringing a water bottle for staying hydrated during the hike. While a water bottle is undoubtedly necessary during hiking, it should not be listed as a breakfast food. Humans would not consider it a correct answer to the same problem.

In addition, we observed that ChatGPT's performance tended to degrade when asked questions related to time estimation. For example, when the model was asked how many days Fred and Linda would be away on their 10-day camping vacation, given that it takes one full day to drive to the White Mountains of New Hampshire and one full day to drive back home, ChatGPT responded with ten days, which is not an accurate answer (a better answer is 9 days). Similarly, when asked to estimate the time required to set up a four-person tent, ChatGPT's answer of 30 min to an hour did not match the range of 5–30 min provided by humans. This discrepancy may be deemed incorrect in a multiple-set question setting, despite being (somewhat) acceptable in a generative question setting. Other instructive details can be obtained from the full log linked earlier.

[1] https://docs.google.com/document/d/1yNrjTOt0imJW5OVajTDNcAJ0PJxmM6B7Dpe3N8YFMD4/edit?usp=sharing.

During the second evaluation session[2], ChatGPT was tested on its ability to apply commonsense reasoning to organize a series of camping activities in the correct order. To achieve this goal, four distinct questions were presented to the model, including "What activities are best to do before it gets dark while camping?" Impressively, ChatGPT provided the correct sequence of activities for all the questions.

In addition to grading the results, involving a human in the evaluation loop allowed us to observe that ChatGPT recognized that the order of activities could be influenced by specific circumstances. For instance, ChatGPT noted that weather conditions and the availability of firewood at the camping site could impact the order of activities before nightfall. Moreover, when asked about sequencing activities before nightfall, the model suggested that campers should be mindful of the campground's quiet hours and avoid making excessive noise during the evening. Probing the model's abilities to handle such context is currently not allowed by single-hop QA paradigms. However, our proposed protocol is flexible enough to include such context when constructing queries.

Overall, ChatGPT exhibited impressive commonsense reasoning abilities in organizing a series of camping activities in the correct order. Including a human in the evaluation loop helped us to assess the model's performance and provided additional valuable insights into the model's strengths and limitations.

### 3.2 Context 2: Vacationing Abroad

The second problem context in the assessment relates to the notion of vacationing abroad. Chloe, who has not taken a vacation in nearly two years, plans to take an entire month off. She intends to spend three weeks traveling with some close friends to visit Europe's most renowned attractions, such as Paris and London. We assess the extent to which the system comprehends Chloe's vacation plans by requesting it to carry out an intent-analysis of Chloe's itinerary elements and offer a rough estimation of the traveling agenda.

The evaluation was conducted in a multi-set question session and a generative question session, respectively[3]. The same set of questions was used in both sessions to compare the performance of ChatGPT on different evaluation tasks. In the multi-set question session, a set of candidate answers was provided per question, and the model was asked to select all the correct options that apply. In contrast, the generative question session allowed ChatGPT to freely generate its response.

In the generative question session, ChatGPT performed reasonably well, but in the multi-set question session, it only correctly answered 6 out of 12 questions.

[2] The log for this session may be found at https://docs.google.com/document/d/1a-CDcijT2an0XiYF-JQ0i2ZvFiUpUB-Xb4wZBUkBVkg/edit?usp=sharing.

[3] The logs for these sessions may be found at https://docs.google.com/document/d/1tLseMBfGVEhdpcm4jGNg_9Dr4ruGihFX9ncY3240k5Y/edit?usp=sharing and https://docs.google.com/document/d/1HWma7MuZkaCeqq6aVmXtBzP9pqudmF7YH1GAl9z2xoc/edit?usp=sharing, respectively.

One example of a discrepancy between the two sessions is the question, "While Chloe likes outdoor activities, she doesn't appreciate them when it's sunny and hot. During her trip to Europe, what should she do during the day?". ChatGPT chose inappropriate candidate answers such as "Get a coffee" and "Have dinner in a new place," but generated more relevant responses in the generative question session, such as visiting museums, shopping at indoor markets, and taking a river tour. Determination of the closeness of an answer by the generative reasoner is based on two methods: the first is a text match (i.e., when an answer matches the text of an option, then it is considered exact or close), but when such a match is not detected, the second method relies on a human in the loop to determine the closeness of the match. This suggests that, despite some suggestions to the contrary that these models are 'general' in their abilities, it may not necessarily be the case that generative models are better at discriminative tasks (such as multiple-choice QA).

Our human-reviewed evaluation indicated that ChatGPT's performance on time-related questions was still not up to par in both assessment sessions. Reasoning about time is a foundational commonsense reasoning skill and is required in order to reason about related commonsense issues like activities and planning [7]. Temporal reasoning is one of the foundational capabilities that researchers [1] believe is necessary in order for machine-based reasoning systems to perform basic tasks or to support humans in those tasks, e.g., resource management, travel planning. For instance, when presented with the question, "Given that Chloe's vacation starts on June 1st and she only has three weeks of vacation, when should her flight depart?", ChatGPT only identified "June 1st" as the correct answer, even though "June 5th" was also a valid option provided in the candidate answer list. The model acknowledged that the other options could be correct if additional information, such as a specific flight departure time or a fixed schedule, was provided. However, in our annotation exercises, we found that humans could easily choose both answers as appropriate without any additional hints.

Furthermore, when presented with the same question without a list of candidate answers in the generative QA session, ChatGPT responded that Chloe should depart on or after June 1st and return on or before June 22nd. While this answer is correct, it still shows that the model struggles with accurately understanding and processing time-related information. In fact, this answer is consistent with June 5th, but the model did not choose it when presented with it in a discriminative setting.

## 4 Discussion

This paper proposed a context-rich evaluation framework where limited human intervention is used for two important purposes: to determine if the response to a query by the model is actually appropriate and along what dimensions (which can be difficult to automate, especially if the query was ambiguous), as well as to incrementally dialog with the machine in order to more robustly evaluate its multi-hop reasoning capabilities. To more effectively evaluate a machine's ability to solve different types of commonsense problem-solving tasks, we recommend that the content of questions be grounded in a theory of commonsense, such as that proposed and refined in [7]. We demonstrated the potential utility of this framework by applying it to the ChatGPT system for assessing its MCS abilities. Although the model's responses are quite impressive, the full use of the evaluation protocol also demonstrates that more work is needed before the model can be said to possess the full gamut of common sense reasoning.

As enterprises and other practitioners (including in healthcare and education) start deploying generative AI technologies like LLMs more frequently in their application stacks, *domain-specific* evaluations of such models could prove critical in ensuring that they are being used in a responsible and trustworthy manner. The protocol proposed in this paper could be adapted for domain-specific evaluations; the only real constraint would be to ensure that the *problem context* in Fig. 1 aligns appropriately with the domain, and the human-in-the-loop evaluator is an individual with sufficient domain expertise.

Beyond partnering with domain experts on such evaluations, in future work, we plan to scale up evaluations significantly and apply the protocol to other LLMs as they are released. We also hypothesize that, by applying the protocol rigorously in multiple sessions, deeper insights could be gleaned on the MCS capabilities and limitations of generative models. By focusing more on a benchmarking process, rather than over-reliance on benchmarking data (which has been found to be susceptible to generalization issues [11]), similar such evaluation sessions could then be conducted and replicated as larger models continue to be released. By releasing the session logs, as we have done in this work, the process also becomes open to analysis and could be refined or modified through community-driven critique.

## References

1. Allen, J.F.: Maintaining knowledge about temporal intervals. Commun. ACM **26**(11), 832–843 (1983). https://doi.org/10.1145/182.358434
2. Ammanabrolu, P., Broniec, W., Mueller, A., Paul, J., Riedl, M.: Toward automated quest generation in text-adventure games. In: Proceedings of the 4th Workshop on Computational Creativity in Language Generation, pp. 1–12. Association for Computational Linguistics, Tokyo, Japan (2019). https://aclanthology.org/2019.ccnlg-1.1
3. Bang, Y., et al.: A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023 (2023)
4. Blagec, K., Dorffner, G., Moradi, M., Samwald, M.: A critical analysis of metrics used for measuring progress in artificial intelligence. arXiv preprint arXiv:2008.02577 (2020)
5. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)

6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423
7. Gordon, A.S., Hobbs, J.R.: A Formal Theory of Commonsense Psychology: How People Think People Think. Cambridge University Press, Cambridge (2017). https://doi.org/10.1017/9781316584705
8. Gunning, D.: Machine common sense concept paper. arXiv preprint arXiv:1810.07528 (2018)
9. Kahneman, D.: Thinking, Fast and Slow. Macmillan (2011). Google-Books-ID: SHvzzuCnuv8C
10. Kejriwal, M., Santos, H., Mulvehill, A.M., McGuinness, D.L.: Designing a strong test for measuring true common-sense reasoning. Nat. Mach. Intell. **4**(4), 318–322 (2022). https://doi.org/10.1038/s42256-022-00478-4
11. Kejriwal, M., Shen, K.: Do fine-tuned commonsense language models really generalize? arXiv preprint arXiv:2011.09159 (2020)
12. Lin, B.Y., et al.: CommonGen: a constrained text generation challenge for generative commonsense reasoning. In: Findings of the Association for Computational Linguistics (EMNLP 2020), pp. 1823–1840. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.165
13. Mitra, A., Banerjee, P., Pal, K.K., Mishra, S., Baral, C.: How additional knowledge can improve natural language commonsense question answering? arXiv preprint arXiv:1909.08855 (2020)
14. Neelakandan, N.: Creating multiple-choice questions: the dos and don'ts (2019). https://elearningindustry.com/creating-multiple-choice-questions
15. Santos, H., Kejriwal, M., Mulvehill, A.M., Forbush, G., McGuinness, D.L., Rivera, A.R.: An experimental study measuring human annotator categorization agreement on commonsense sentences. Exp. Res. **2**, e19 (2021)
16. Santos, H., Shen, K., Mulvehill, A.M., Razeghi, Y., McGuinness, D.L., Kejriwal, M.: A theoretically grounded benchmark for evaluating machine commonsense (2022). https://doi.org/10.48550/arXiv.2203.12184
17. Shen, K., Kejriwal, M.: An experimental study measuring the generalization of fine-tuned language representation models across commonsense reasoning benchmarks. Expert Syst. e13243 (2023)