Thinking Out-of-the-Box: A Comparative Investigation of Human and LLMs in Creative Problem-Solving

Anonymous Authors[1]

Abstract

We explore the creative problem-solving capabilities of modern LLMs in a novel constrained setting. To this end, we create MACGYVER, an automatically generated dataset consisting of over 1,600 real-world problems deliberately designed to trigger *innovative usage of objects* and necessitate *out-of-the-box thinking*. We then present our collection to both LLMs and humans to compare and contrast their problem-solving abilities. MACGYVER is challenging for both groups, but in unique and complementary ways. For instance, humans excel in tasks they are familiar with but struggle with domain-specific knowledge, leading to a higher variance. In contrast, LLMs, exposed to a variety of specialized knowledge, attempt broader problems but fail by proposing physically-infeasible actions. This work **(1)** introduces a fresh arena for intelligent agents focusing on intricate aspects of physical reasoning, planning, and unconventional thinking, which supplements the existing spectrum of machine intelligence; and **(2)** provides insight into the constrained problem-solving capabilities of both humans and AI.

1. Introduction

Creativity has long been considered the driving force behind modern civilization, and one of the hallmarks of human intelligence (Guilford, 1967b; Hennessey, 1995). As large language models (LLMs) become increasingly powerful, researchers investigate their reasoning ability in problem-solving tasks (Yao et al., 2022; Brahman et al., 2023) and their capacity for creativity as demonstrated by expressing humor and generating artistic content (Mittal et al., 2022; Hessel et al., 2023; Ramesh et al., 2022; Chakrabarty et al., 2022; Tian et al., 2023). However, everyday activities that involve creative thinking have not been studied to the same extent. In this work, we contribute a benchmark for creative problem solving, hoping to critically assess modern LLMs when it comes to 'thinking out-of-the-box'.

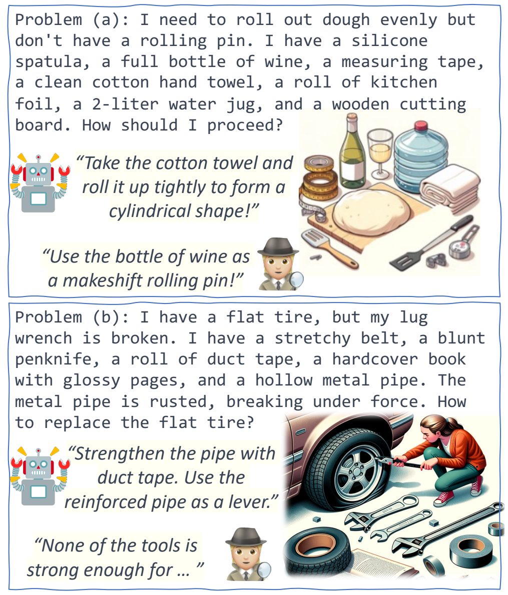


Figure 1. Examples of the problems in our MACGYVER dataset with the GPT-4 and human answers (continued in Figure 8). Pictures, drawn by DALL·E 3, are solely for illustration purposes and may not accurately reflect the text. In our experiment, all inputs to human and LLMs are natural language texts.

To bridge this gap, we curate MACGYVER, ***a novel unconventional problem-solving dataset*** consisting of 1,683 sets of verbal problems that require human-like creativity in the realm of physical reasoning. Drawing inspiration from the cognitive science literature (Duncker & Lees, 1945), we collect problem scenarios that deliberately push against *functional fixedness*—a cognitive bias that limits an agent from employing familiar tools in innovative ways. Notably, leveraging the *generative* strength of LLMs and the *verification* strength of humans, we design a novel and labor-efficient pipeline to collect progressively more challenging scenarios (Section 2). These scenarios are verified by humans as requiring unconventional usage of objects to find a solution. For example, solving problem (a) in Figure 1 requires

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

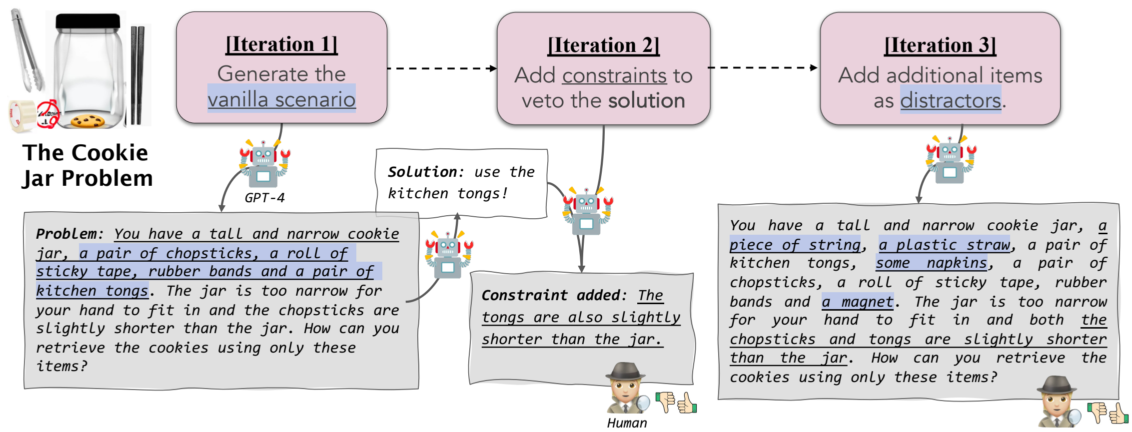


Figure 2. Progressive problem refinement with GPT-4. Starting from a vanilla version (*i.e.,* Iteration 1), we carefully design refinement steps that gradually increase the problem's complexity by adding specific object properties as constraints to veto a previous solution (*i.e.,* Iteration 2), and adding distracting objects that are (likely) not involved in the solution the problem (*i.e.,* Iteration 3). After that, human verifiers judge the quality of refined problems.

using the wine bottle as a makeshift rolling pin.[1] Each problem in our dataset is paired with at least one human-provided or verified solution. To the best of our knowledge, MacGyver is the first dataset of unconventional everyday problems requiring two key elements of creativity (Guilford, 1967a): ***divergent*** thinking (to come up with creative or unconventional usage of objects) and ***convergent*** thinking (to accomplish a goal efficiently).

Next, we use the resulting dataset as a ***benchmark*** to evaluate the creative problem-solving abilities of both human participants and recent LLMs, including GPT-3.5, GPT-4, PaLM2, Claude2, and Llama2 (OpenAI, 2022; 2023; Anil et al., 2023; Touvron et al., 2023; Anthropic, 2023). Our results in Section 3 reveal a substantial gap between most LMs and human. While the best performing LM, GPT-4, complements the capability of an arbitrary human under certain domain-specific settings (*e.g., fixing a hole on the wall*), humans' collective wisdom is so far still invincible. Additionally, LLMs struggle to identify unsolvable problems and either exhibit misleading helpfulness or are ultraconservative in inappropriate cases. Finally, in Section 4, we present detailed comparison between human and machine, demonstrating their complementary strengths.

2. MacGyver Dataset

LLMs have demonstrated utility for idea generation (Girotra et al., 2023). Therefore, instead of asking humans to come up with thousands of constrained scenarios from scratch, we design a progressive refinement pipeline to explore LLMs' potential to generate problem settings quickly and at scale (§2.1). Human annotators then verify that each problem is concrete and requires creativity (Section 2.2). Each instance in our dataset includes a constrained problem setting paired with at least one human-provided or verified solution (Section 2.2, §H.2).

2.1. Progressive Problem Refinement for Dataset Creation

Figure 2 illustrates of our problem collection pipeline, showing how we combine human and machine inputs. Specifically, we propose a progressive problem refinement approach that gradually increases problem complexity by 1) adding specific object properties (*e.g.,* material, size, etc.) as constraints to eliminate a previous solution and 2) adding distracting objects that are not involved in the solution. From a cognitive perspective on problem-solving (Knoblock, 1991), the first refinement step removes the most straightforward solution path, while the second step further complicates the problem by adding branches to the search space. We implement this pipeline through a dialogue interaction with GPT-4. Human assessment results (detailed in Appendix H.3) confirm that both steps within the progressive refinement approach pose additional challenges to LLMs, and after the two iterations, the original problem requires more creativity and becomes more challenging.

2.2. Human Verification Process

After the refinement process, we involve human verifiers to judge if the final versions of the problems **1)** are solvable, unsolvable, or need more clarification (*e.g.,* the setup is vague, which will be discarded), and **2)** for those solvable, whether solving them efficiently requires creative thinking (*i.e.,* using objects to achieve goals they were *not* originally designed for —unconventional usage). Each problem is annotated by three human verifiers, with average inter-

[1] If the problem is unsolvable given the presented tools and constraints (problem b in Figure 1), we expect the agent to identify such infeasibility and provide a short justification.

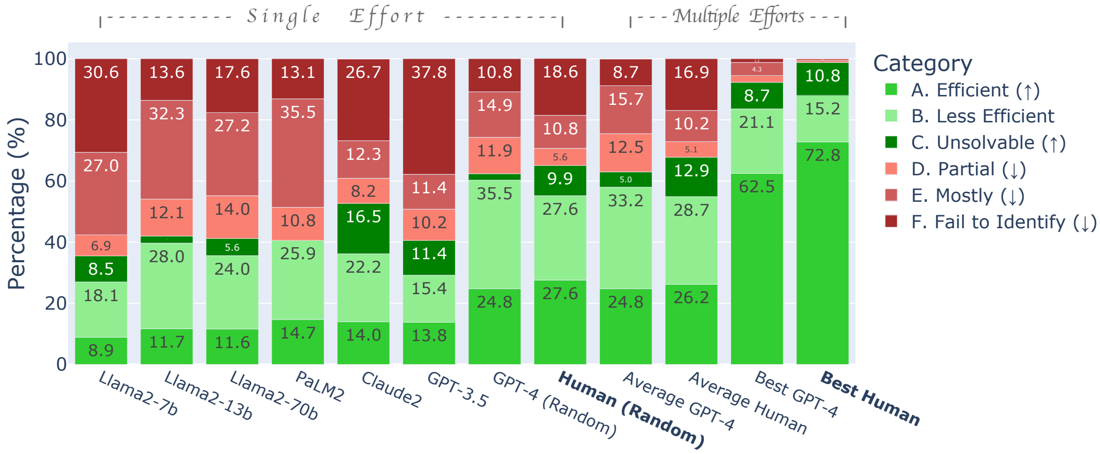


Figure 3. **Left**: Benchmark results of seven LLMs and human with a single effort. For human participants, since there is no single participant who worked on all problems, we take a random response from each problem. We color-code the three categories indicating fine-grained aspects of correctness or falseness. **Right**: Comparison between GPT-4 and human where we evaluated multiple solutions per problem. The best performance, which can be viewed as an upper bound, is computed by taking the individual best answer (out of 6) for each problem. The actual numbers are reported in Table 3 in Appendix B.2.

Problem (All)	**Solvable**	**Unsolvable**	**Total**
Count	1,306	377	1,683
Percentage	77.6%	22.4%	100%

Problem (Solvable Subset)	**Unconv.**	**Conv.**	**Total**
Count	1,073	233	1,306
Percentage	82.2%	17.8%	100.0%

Table 1. Statistics of the entire MacGyver dataset (top), and solvable problems that require unconventional use of tools (bottom).

annotator agreement (IAA, measured by Cohen's Kappa) of 0.67 and 0.77 for tasks **1)** and **2)**, respectively. Finally, we pair each problem with a gold answer. For the solvable subset, it is a step-by-step feasible solution. For the unsolvable subset, it is an explanation why the stated goal cannot be achieved (detailed in §H.2).

In total, we created 1,683 problems, with a detailed breakdown in Table 1. Of those, 78% are solvable and 22% are unsolvable. Another 7% of all problems were discarded after being annotated by at least one annotator to be ambiguous or contradictory. For solvable problems, 82% require using tools in an innovative or unconventional manner.

Finally, we take measures to ensure the collected problems are *diverse, comprehensive, and free of repetitive patterns* in § H.4. In total, over 3,800 unique tools are included in MacGyver dataset.

3. Benchmarking Humans and LLMs

A natural follow-up question is how well modern LLMs perform on this task, as compared to humans. We thus evaluate the performance of several recent LLMs (*i.e.,* PaLM2, Claude2, Llama2, GPT-3.5 and GPT-4) on a representative sample of the entire MacGyver dataset which contains 323 problems. In addition, we gauge the capability of average humans on the same set of tasks.

Next, a different set of human experts were asked to evaluate if a presented answer is correct by selecting one out of six fine-grained categories: **A (or B)** correctly giving a feasible and efficient (or less efficient) solution to a solvable problem; **C** correctly identifying an unsolvable problem and giving the right justification; **D** giving a partially incorrect answer; **E** giving a mostly or entirely wrong answer; and **F** failing to identify the correct solvability status.

3.1. Benchmark Results

We report the benchmark results in Figure 3. Category **A**, **B**, and **C** are the three aspects of correct responses, while the remaining **D**, **E**, and **F** are aspects of the wrong ones. At a glance, despite varying in their characteristics, all of the benchmarked LLMs lag behind the performance of humans.

Performance with Single Effort. We first list the LLMs' performances with their ***single best answers*** on left of Figure 3.For human participants, there is no single person who approached all problems. Therefore, to simulate ***an arbitrary person's individual*** performance, we take a random response from each problem.

We see that most recent LLMs achieve a mere 35% to 42% chance of success. Although GPT-4 and Claude2 stand out among the tested LLMs, their best attempts still underperform an arbitrary average person with total correct rate of 65.1% (sum of category **A**, **B** and **C**). Different families of LLMs exhibit dissimilar behaviors. For example, PaLM2 and GPT-4 are overly verbose and often suggest solutions to problems that are inherently unsolvable. In contrast, Llama2-7b, Claude2, and GPT-3.5 are more conservative and fail to realize a constrained problem can still be solvable (reflected in their high numbers in category **F**). Comparing the three variants of Llama2, we find that the larger models (13b, 70b) excel in correctly identifying solvability (category **F**). The smaller model (7b) is more subject to falsely recognizing a constrained problem as unsolvable. Beyond this, however, it appears that scale alone does not significantly unleash any creative problem-solving capabilities.

Performance with Multiple Efforts. We collect multiple solutions per problem from both GPT-4 and humans, calculating their ***average*** and ***best*** performance. The best performance, an upper bound, is the highest score from individual answers per problem. Additionally, we assess the ***majority*** performance based on binary correctness, finding humans at 79.3% accuracy, above GPT-4's 73.3%.

On average, human participants are slightly worse than GPT-4 in coming up with a correct solution (especially inefficient ones in category **B**), possibly due to functional fixedness. However, they generally surpass GPT-4, particularly as GPT-4 often fails to recognize unsolvable problems. The best human answer, reflecting collective wisdom, nearly achieves perfection.

Finally, human seem to struggle with certain challenges (category **F**). We hypothesize that an individual, lacking expertise in all life domains, may not surpass a comprehensive LLM like GPT-4, which is trained on massive amount of data and a wide variety of tasks. Yet, collectively, human intelligence, enriched by diverse expertise and insights, outstrips LLMs. We explore this further in the next section.

4. Comparing GPT-4 with Humans

4.1. Humans have higher variance than LLMs.

We plot the kernel density estimate (KDE) of individual human and GPT-4 responses in Figure 4. We can see that humans either approach a problem perfectly or fail totally. Namely, once humans understand the task and acquire the relevant knowledge, they can always propose a feasible and often the most efficient solution. On the contrary, GPT-4 responses fall more into the middle (mostly/partially wrong, or inefficient), owning to its ability to aggregate information from a wide range of sources it has been trained on. However, GPT-4 is sometimes ignorant of tool affordances or consequences of its proposed actions, lacking the depth of understanding that humans possess (see more detailed error analysis in §D.1).

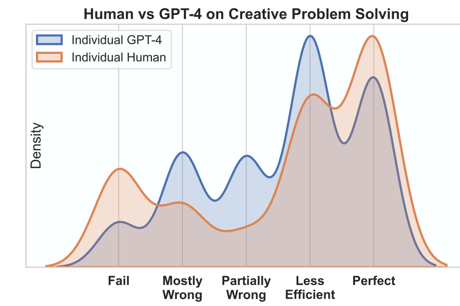


Figure 4. The kernel density estimate of individual human and GPT-4 answers.

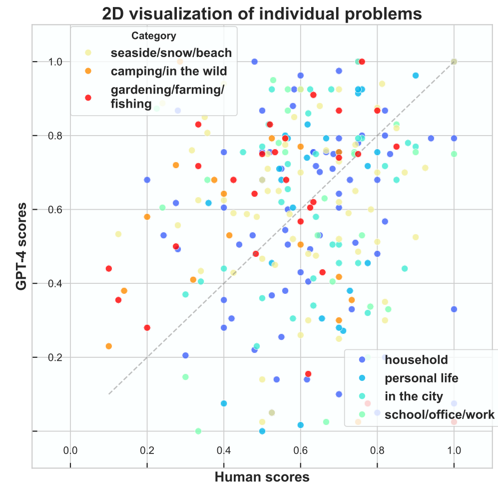


Figure 5. 2D visualization of human (x-axis) and GPT-4 (y-axis) performance on individual problems. Each dot represents a problem, with its color representing seven different categories. Humans are better at solving problems that they are familiar with (*e.g.,* household), than those requiring domain-specific knowledge (*e.g.*, gardening/farming/fishing).

4.2. Humans possess better general everyday knowledge, but less domain-specifically.

Next, we visualize the capability of humans and GPT-4 on individual problems in a 2D plot (Figure 5). Accordingly, we convert categorical labels into numerical scores ranging from 0 (Fail) to 1 (Perfect), and take the average score across solutions. We also plot the diagonal line: the farther away a point is from this, the larger the gap between human and GPT-4 performance.

We find that humans are better at solving tasks in categories likely to be familiar to them, such as *household* and *personal life*. For those requiring domain-specific knowledge such as *gardening/farming/fishing*, GPT-4 performs better. The same holds when we manually inspect the outliers: those few problems that belongs to everyday categories yet humans are poor at. Unsurprisingly, they are problems such as demonstrating the concept of refraction without a prism (category: school), and making a sundial (category: beach), which an average person might have little experience with. Refer to §B.1 for examples and other comparisons.

Overall, the different creative strengths of humans and AI systems suggests that the most effective solutions to tasks requiring thinking "out-of-the-box" might arise from a *collaborative approach* leveraging the strengths of both parties.

References

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Anthropic. Model card and evaluations for claude models. 2023. URL https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf.

Aroca-Ouellette, S., Paik, C., Roncone, A., and Kann, K. Prost: Physical reasoning about objects through space and time. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4597–4608, 2021.

Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L., and Girshick, R. Phyre: A new benchmark for physical reasoning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/4191ef5f6c1576762869ac49281130c9-Paper.pdf.

Bisk, Y., Zellers, R., Gao, J., Choi, Y., et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Brahman, F., Bhagavatula, C., Pyatkin, V., Hwang, J. D., Li, X. L., Arai, H. J., Sanyal, S., Sakaguchi, K., Ren, X., and Choi, Y. Plasma: Making small language models better procedural knowledge models for (counterfactual) planning. *arXiv preprint arXiv:2305.19472*, 2023.

Chakrabarty, T., Padmakumar, V., and He, H. Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6848–6863, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.460. URL https://aclanthology.org/2022.emnlp-main.460.

Collins, K. M., Wong, C., Feng, J., Wei, M., and Tenenbaum, J. B. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. *arXiv preprint arXiv:2205.05718*, 2022.

Duncker, K. and Lees, L. S. On problem-solving. *Psychological monographs*, 58(5):i, 1945.

Girotra, K., Meincke, L., Terwiesch, C., and Ulrich, K. T. Ideas are dimes a dozen: Large language models for idea generation in innovation. *Available at SSRN 4526071*, 2023.

Guilford, J. P. Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior*, 1(1):3–14, 1967a.

Guilford, J. P. The nature of human intelligence. 1967b.

Hennessey, B. A. Social, environmental, and developmental issues and creativity. *Educational Psychology Review*, 7: 163–183, 1995.

Hessel, J., Marasovic, A., Hwang, J. D., Lee, L., Da, J., Zellers, R., Mankoff, R., and Choi, Y. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 688–714, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.41. URL https://aclanthology.org/2023.acl-long.41.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Hong, Y., Yi, L., Tenenbaum, J. B., Torralba, A., and Gan, C. PTR: A benchmark for part-based conceptual, relational, and physical reasoning. *CoRR*, abs/2112.05136, 2021. URL https://arxiv.org/abs/2112.05136.

Kaufman, J. C. and Beghetto, R. A. Beyond big and little: The four c model of creativity. *Review of general psychology*, 13(1):1–12, 2009.

Knoblock, C. A. Search reduction in hierarchical problem solving. In *AAAI*, volume 91, pp. 686–691. Citeseer, 1991.

Koivisto, M. and Grassini, S. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, 13(1):13601, 2023.

Liu, A., Wu, Z., Michael, J., Suhr, A., West, P., Koller, A., Swayamdipta, S., Smith, N., and Choi, Y. We're afraid language models aren't modeling ambiguity. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 790–807, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.51.

Mittal, A., Tian, Y., and Peng, N. AmbiPun: Generating humorous puns with ambiguous context. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1053–1062, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.77. URL https://aclanthology.org/2022.naacl-main.77.

OpenAI. Introducing chatgpt. 2022. URL https://openai.com/blog/chatgpt.

OpenAI. Gpt-4 technical report, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Reimers, N. and Gurevych, I. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. URL https://arxiv.org/abs/2004.09813.

Tian, Y., Narayan-Chen, A., Oraby, S., Cervone, A., Sigurdsson, G., Tao, C., Zhao, W., Chung, T., Huang, J., and Peng, N. Unsupervised melody-to-lyrics generation. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9235–9254, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.513. URL https://aclanthology.org/2023.acl-long.513.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wang, Y. R., Duan, J., Fox, D., and Srinivasa, S. Newton: Are large language models capable of physical reasoning? *arXiv preprint arXiv:2310.07018*, 2023.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Zellers, R., Bisk, Y., Schwartz, R., and Choi, Y. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 93–104, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1009. URL https://aclanthology.org/D18-1009.

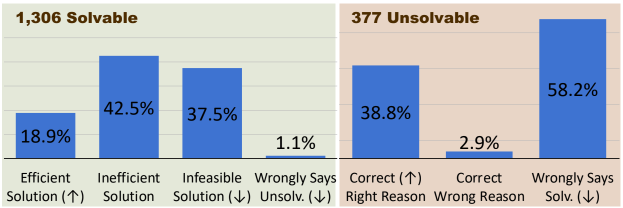


Figure 6. Left: Human-evaluated GPT-4 performance on all 1,306 problems from the MACGYVER that humans think are **solvable**. Right: GPT-4 performance on all 377 problems that humans think are **unsolvable**. *Correct for the right reason* means that the LLM correctly identifies the problem is unsolvable, and gives the right justification. *Correct for the wrong reason* means that it correctly identifies the problem is unsolvable, but gives an incorrect justification.

A. Assessing the Task Difficulty

To gauge the challenge of our task posed to the most recent LLMs, we evaluate the zero-shot performance of GPT-4 (OpenAI, 2023). Nevertheless, existing automatic evaluations fall short to assess the efficacy of a presented solution. Therefore, we recruit human annotators to evaluate the quality of the GPT-4's answers on the ***entire*** MACGYVER.

Assessment Setup. For a solvable problem, human annotators are asked to judge if the presented solution is **1.1** *feasible and efficient*[2], **1.2** *feasible yet inefficient*, or **1.3** *infeasible*. The machine-generated answer may also wrongly assume the problem is unsolvable and gives a wrong justification (**1.4**). For an unsolvable problem, they need to judge if the presented answer **2.1** *correctly identifies the problem as unsolvable*, and **2.2** *gives the right justification*. Similarly, the answer may also wrongly assume the problem is solvable and give a wrong solution (**2.3**).

GPT-4 Performance. We report the performance on the solvable and unsolvable subset in Figure 6. Our preliminary findings indicate that, **firstly**, LLMs as strong as GPT-4 still exhibit limitations in solving unconventional problems, with only 18.9% likelihood of providing an efficient solution, while 37.5% likelihood of providing an infeasible solution. Analysis in the later section (Appendix D) shows that one common mistake is it failing to realize the consequences of actions and tool affordances in the given context (*e.g.,* proposing to use chopsticks to lift up the egg yolk). **Secondly**, GPT-4 displays overconfidence, often suggesting solutions to problems that are inherently unsolvable. This could be partially due to GPT-4 being trained with RLHF (Ouyang et al., 2022), maximizing its helpfulness. Moreover, the model struggles to discern whether a problem description is sufficiently concrete for resolution or too ambiguous, necessitating additional context (Liu et al., 2023).

[2] A solution is considered efficient if it has no redundant or unnecessary steps, and it is unlikely that the problem can be solved with less labor or using fewer steps.

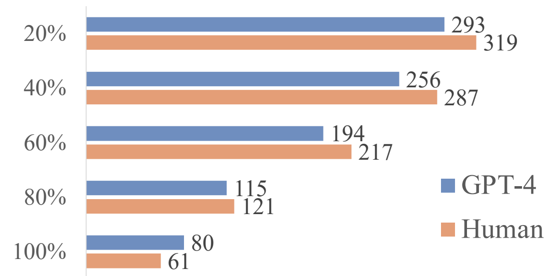


Figure 7. Number of problems (out of 323) that at least 20%, 40%, 60%, 80%, 100% human participants (or GPT-4) answer correctly.

B. Additional Results

B.1. Comparing GPT-4 with Humans

What percentage of individual humans outperform AI? Figure 7 compares human and machine by showing the number of problems (out of 323) that at least 20%, 40%, 60%, 80%, and 100% human participants (or GPT-4) answer correctly. Given the unique strengths and knowledge scopes of different individuals, it is less likely that all human participants can answer the same problem correctly. However, there is a higher chance where at least 60% human participants know the answer.

Complementary capabilities of human and AI. Figure 8 presents two examples showing the complementary capabilities of human and AI in creative problem solving. In problem (c), human participants find a more efficient solution to heat the pizza than GPT-4. In problem (d) requiring domain knowledge gardening, humans fail to solve this highly-specialized task, whereas the LLM has equipped itself with such domain knowledge during massive pre-training.

What tools are human more proficient at? Recall that in Section 4.2 we convert the categorical labels into numerical scores ranging from 0 (Fail) to 1 (Perfect) to conduct problem-wise analysis. Similarly, we conduct object-wise analysis by first parsing the tools presented in each problem, and then calculating the same numerical scores for each tool. Note that we opt to parse all the tools presented in the problem setting instead of those actually used in a proposed solution, because being able to reason about the potential usage of presented tools and conclude to *not* use a possible tool is also an keystone towards intelligence. We identify several tools that humans and GPT-4 attempt most differently and report them in Table 2. For example, humans are more proficient at attempting magnifying glass, rocks, calculators, knifes, etc., whereas AIs are better attempting

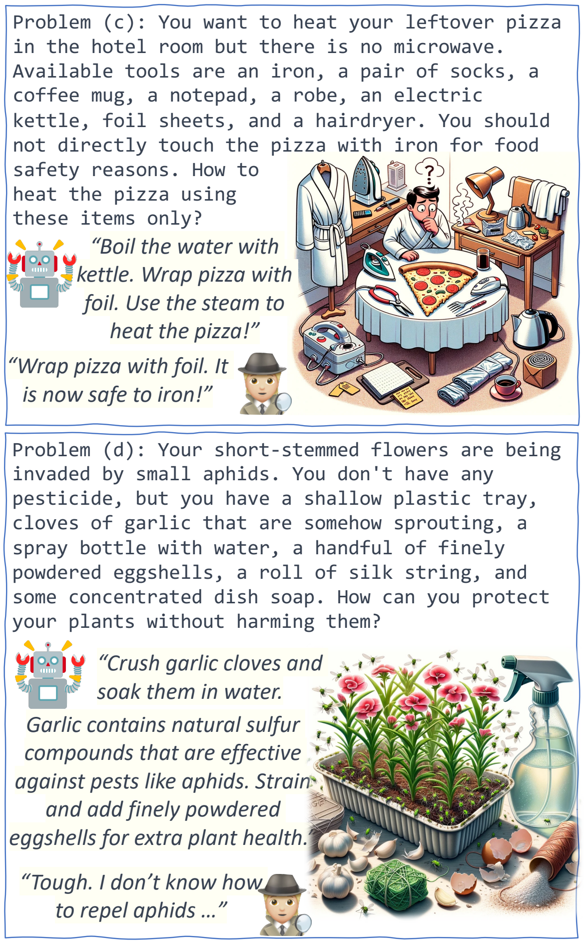


Figure 8. Detailed examples showing the complementary capabilities of human and GPT-4. In problem (c), human participants find a more efficient solution to heat the pizza than AI. In problem (d), humans fail to solve this highly-specialized task to repel aphids, whereas the LLM has equipped itself with domain knowledge on gardening during massive pre-training.

	Object	Human-AI Difference
A. human>AI	magnifying glass	0.602
	rock	0.447
	calculator	0.405
	kitchen knife	0.386
	hair tie	0.359
	paper cup	0.292
	zip ties	0.283
	pen	0.281
	kettle	0.273
	old t-shirt	0.252
	sunscreen	0.25
B. human<AI	mirror	-0.314
	gardening gloves	-0.311
	scarf	-0.307
	tablecloth	-0.289
	clothespins	-0.253

Table 2. Tools that human are more proficient at leveraging or deciding to not leverage than AI (GPT-4 in our case), and vice versa.

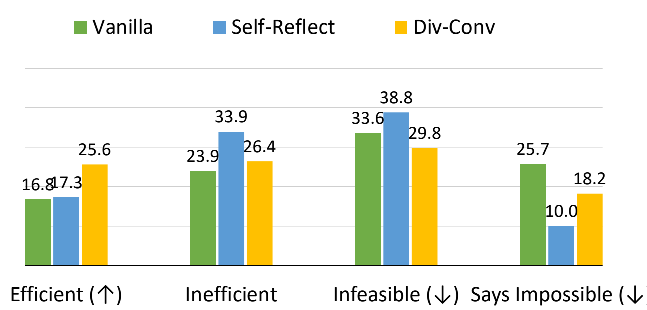


Figure 9. Results of different prompting strategies with Claude2. We compare 1) vanilla prompting, 2) iterative step-wise reflection (reflect), and 3) divergent-convergent thinking (div-conv).

mirrors, gloves, and scarves. In general, there are more tools humans are proficient at.

B.2. Benchmark Results

We report the benchmark results in Table 3. Category **A**, **B**, and **C** are the three aspects of correct responses, while the remaining **D**, **E**, and **F** are aspects of the wrong ones. At a glance, despite varying in their characteristics, all of the benchmarked LLMs lag behind the performance of humans.

B.3. Enhancing LLMs' Problem Solving

Results with Claude2 and Llama2 We report the performance of the standard, zero-shot prompting and two proposed improvements for Claude2 and Llama2-13b in Figure 9 and Figure 10. Different from GPT-4 (shown in Figure 12), the `self-reflection` strategy does not help any of these two models to reduce infeasible answers. When prompted to reflect on its previous answer, Llama2 always claims that its original answer is mistaken and attempts to correct itself blindly. We hypothesize that these two LLMs are weaker than GPT-4 and lack the inherent ability to faithfully conduct complicated physical reasoning. On the other hand, we see that `Divergent-Convergent Thinking` is beneficial for all LLMs across all dimensions.

C. The Four-C Creativity Model

Kaufman & Beghetto (2009) propose the Four-C model (Table 5), categorizing human creative activities into `Mini-C`: developmental creativity in the learning process, `Little-C`: everyday innovation that ordinary people have knowledge of and engage with (such as removing wrinkles on a shirt without possession of an iron) , `Pro-C`: profes-

	Correct (%)			Correct in Total (↑)	Wrong (%)			Wrong in Total (↓)
	A. Efficient	B. Less Efficient	C. Unsolvable		D. Partial	E. Mostly	F. Fail to Identify	
				Single Effort				
Llama2-7b	8.9	18.1	8.5	35.5	6.9	27	30.6	64.5
Llama2-13b	11.7	28	2.3	42.0	12.1	32.3	13.6	58.0
Llama2-70b	11.6	24	5.6	41.2	14.0	27.2	17.6	58.8
PaLM2	14.7	25.9	0.0	40.6	10.8	35.5	13.1	59.4
Claude2	14.0	22.2	16.5	52.7	8.2	12.3	26.7	47.2
GPT-3.5	13.8	15.4	11.4	40.6	10.2	11.4	37.8	59.4
GPT-4 (Random)	24.8	35.5	2.1	62.4	11.9	14.9	10.8	37.6
Human (Random)	27.6	27.6	9.9	**65.1**	5.6	10.8	18.6	**35.0**
				Multiple Efforts				
Average GPT-4	24.8	33.2	5.0	63.0	12.5	15.7	8.7	36.9
Average Human	26.2	28.7	12.9	67.8	5.1	10.2	16.9	32.2
Best GPT-4	62.5	21.1	8.7	92.3	2.2	4.3	1.2	7.7
Best Human	72.8	15.2	10.8	**98.8**	0.6	0.6	0.0	**1.2**

Table 3. **Top**: Benchmark results of seven LLMs and human with a single effort. For human participants, there is no single participant who worked on all problems. So we take a random response from each problem. **Bottom**: Comparison between GPT-4 and human where we evaluated multiple solutions per problem. The best performance, which can be viewed as an upper bound, is computed by taking the individual best answer (out of 4) for each problem. We use boldface to denote the best performance and underline to denote the second best.

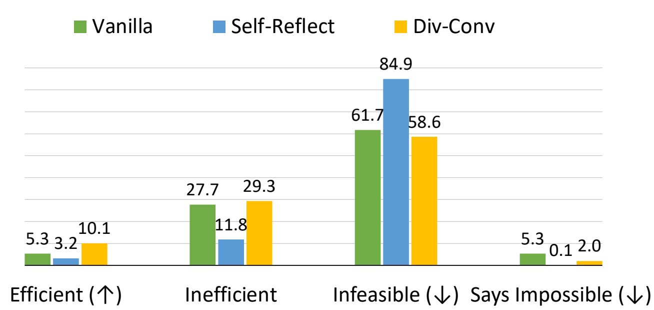


Figure 10. Results of different prompting strategies with Llama2-13b. We compare 1) vanilla prompting, 2) iterative step-wise reflection (reflect), and 3) divergent-convergent thinking (div-conv).

sional expertise such as writing poems or painting artwork, and `Big-C`: highly eminent innovation that few people engage with.

D. Enhancing LLMs' Problem Solving

Here, we investigate whether different prompting strategies can enhance the problem-solving abilities of existing LLMs. In Appendix D.1, we conduct a detailed error analysis on GPT-4, showing it is weakest at identifying the correct tool affordance and physical feasibility. In Appendix D.2, we propose two new prompting strategies that effectively reduce its mistakes.

D.1. Error Analysis for GPT-4

To better understand the limitations of LLMs and provide insight for potential improvement, we manually analyze 200 solutions generated by GPT-4 marked as infeasible by human annotators. We identified five common failure modes in Table 4.

We find that GPT-4 is highly prone to proposing ***physically infeasible, unwanted, or wrong actions***. In Table 4, error type *(1) wrong tool usage* accounts for ∼half of all the errors made (42.4%), followed by *(2) not achieving the goal* (17.7%). It is crucial to highlight that ***LLMs act in a fictional setting***, failing to realize the consequences of their proposed actions and the affordances of tools in the given unconventional context. While one can argue that LLMs lack direct interaction with the physical world, the human solvers similarly contemplate the same task purely in their minds, without any visual or physical cues. We also observe ***two types of hallucination***: *(3) using unavailable tools* and *(5) unfaithful to constraints*, which account for 16.9% + 9.5% = 26.4% of all the errors made.

D.2. Improving LLMs via Prompting

The common error types in Table 4 motivates us to explore techniques to enhance LLMs' problem solving abilities. Specifically, we explore two prompting strategies as illustrated in Figure 11:

[topsep=0pt, itemsep=-2pt, leftmargin=*]**Iterative Step-Wise Reflection** : A self-reflection-based strategy. After the LLM generates an initial solution, we prompt it to *verify* if each step is physically feasible and afforded. Subsequently, it modifies the original solution iteratively until no more modifications are needed. **Divergent-Convergent Thinking**: A cognitive-science-inspired strategy. The LLM is prompted to first enumerate the affordance of each

Error Description	Example	Freq.
(1) Wrong tool usage. Using tools in ways that are physically infeasible or not afforded	Using the stapler to staple the duct tape on top of broken glasses.	42.4%
(2) Not achieving the goal. The proposed approach contains unnecessary or wrong steps towards the stated goal	To save space when packing, use the scissors to cut the comforter into smaller pieces.	17.7%
(3) Using unavailable tools.	-	16.9%
(4) Wrong spatial understanding	Putting the shoe box inside the empty DVD case.	10.8%
(5) Unfaithful to constraints. Ignoring constraints added to a tool or a situation	-	9.5%

Table 4. Categories of common errors made by GPT-4. It is highly prone to coming up with actions that are physically infeasible, unnecessary, or wrong. An erroneous solution may have more than one type of mistake.

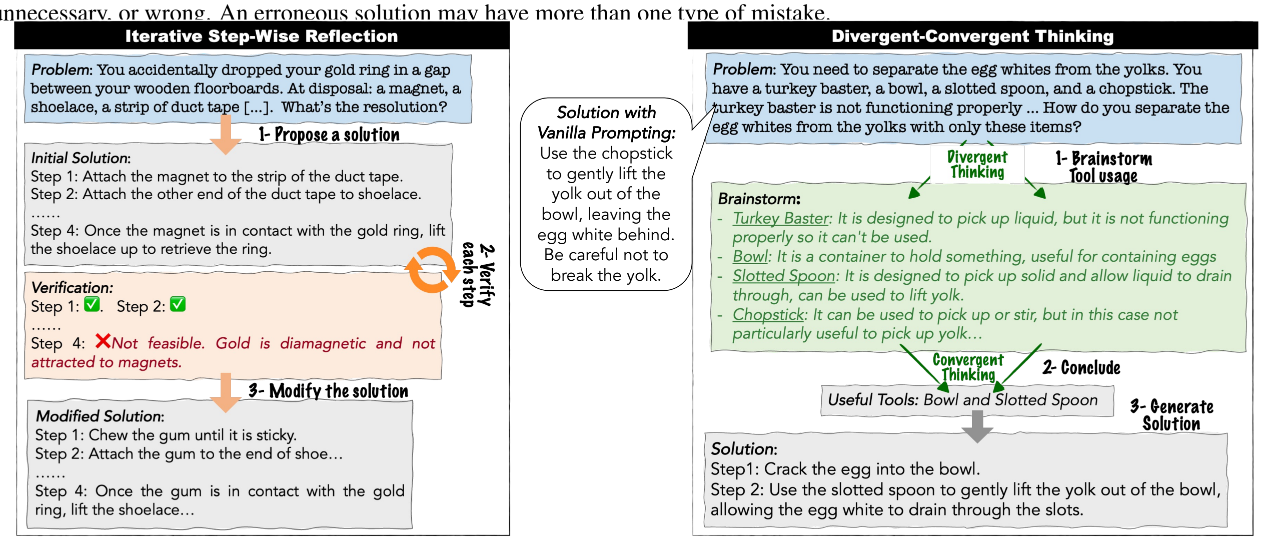


Figure 11. Proposed prompting methods: iterative step-wise reflection (left), divergent-convergent thinking (right).

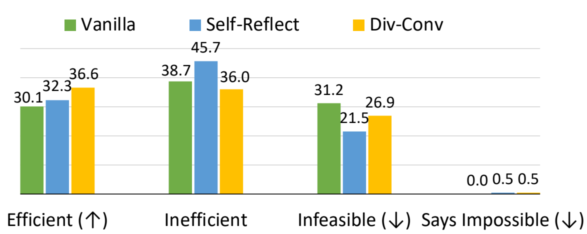


Figure 12. Results of different prompting strategies with GPT-4 in a zero-shot fashion: 1) vanilla prompting, 2) iterative step-wise reflection (self-reflect), and 3) divergent-convergent thinking (div-conv).

object (*i.e.,* divergent thinking) and conclude whether they are useful, followed by generating the steps towards the goal (*i.e.,* convergent thinking).

We implement both prompting strategies with GPT-4, Claude2, and Llama2-13b on 180 randomly-sampled solvable problems that do not overlap with those used in Appendix D.1. The performance of the standard prompting and two proposed improvements for GPT-4 (and the remaining two LLMs) are shown in Figure 12 (and Appendix B.3).

For GPT-4, both proposed prompting methods contribute to a reduction in infeasible solutions. Intuitively, `Self-Reflect`, which is designed to verify the feasibility of steps, has a larger improvement in reducing infeasible solutions (9.7% vs 4.3% drop); while `Div-Conv Thinking`, which is designed for better preparation before generating the solution, is more helpful in generating efficient solutions (6.5% vs 2.2% gain). Comparing all three LLMs, `Div-Conv Thinking` is shown to be beneficial for all, both in terms of efficiency and feasibility, but Claude2 and Llama2's performances do not improve with `Self-Reflect`. Such a finding implies that, smaller models so far still lack the inherent ability to self-reflect and reason about physical consequences which GPT-4 is capable of.

E. Related Work

Creativity Theory Guilford (1967a) defines a meaningful creative process as an interplay between spontaneous (divergent, to come up with novel ideas) and controlled (convergent, to satisfy the demand of the task) modes of thinking. Kaufman & Beghetto (2009) categorize human creative activities into four dimensions (Table 5), ranging from ev-

Model	Description	Example
`Mini-C`	Developmental achievement in the learning process.	A pupil applying a strategy learned in a math class into her science project.
`Little-C`	Everyday innovation that ordinary people engage with.	Removing wrinkles on a shirt without possession of an iron.
`Pro-C`	Professional expertise	Writing poems or stories that receive professional recognition.
`Big-C`	Legendary innovation that redirect a field.	Albert Einstein arriving at general relativity.

Table 5. The Four-C model of creativity.

eryday innovation that ordinary people have knowledge of (*e.g.,* removing wrinkles on a shirt without possession of an iron) to highly eminent innovation that few people engage with.

In the AI-related creativity community, everyday innovation which better reflects the activities that most people may engage in, is under-explored possibly due to the lack of a sizable dataset. For example, recent work (Koivisto & Grassini, 2023) study problems with four objects: rope, box, pencil, and candle. We bridge this gap by contributing a dataset with 1,600 everyday problems. collins2022structured In the AI-related creativity community, everyday innovation which better reflects the activities that most people may engage in, is under-explored possibly due to the lack of a sizable dataset. For example, recent work (Koivisto & Grassini, 2023) study problems with four objects: rope, box, pencil, and candle. We bridge this gap by contributing a dataset with 1,600 everyday problems.

Cognitive Bias *Functional fixedness* is a cognitive bias limiting our ability to use familiar objects in novel ways. For example, struggling to see a chair as anything other than a seat exemplifies this. These biases subtly impact our daily decisions, often unconsciously. Over 82% of the solvable problems in MacGyver require using tools unconventionally to bypass such a bias. A similar work to ours (Collins et al., 2022) explored LLMs' problem-solving ability in out-of-distribution reasoning tasks.

Machine Physical Reasoning Previous research such as Hong et al. (2021) and Bakhtin et al. (2019) investigated physical reasoning in visual contexts. In the realm of language-based physical reasoning, prior studies primarily focused on understanding physical concepts and attributes of various objects, such as PROST (Aroca-Ouellette et al., 2021), and NEWTON (Wang et al., 2023). Relatedly, SWAG (Zellers et al., 2018) introduced the task of grounded commonsense inference about physical situations. PIQA (Bisk et al., 2020), which tests machines' physical commonsense reasoning ability is most similar. While proficiency in addressing problems in MacGyver involves all the above abilities, our emphasis extends beyond. We focus on unconventional tool usage, reasoning over the affordance of tools and ruling out unnecessary ones, and how individual objects can be used in combination to achieve a complex goal.

F. Discussion and Conclusion

Significance of Work We propose a new playground and the accompanying MacGyver dataset for creative problem solving, which covers a broad range of topics for *everyday innovation*, such as household, training, and outdoor sports, which is *orthogonal* to the existing areas of reasoning and creativity, and adds to the spectrum of machine intelligence.

The area of daily innovation, or "little-c" according to the creativity theory (Table 5), is a *stand-alone type* of creativity and better reflects the creative activities that normal people engage with, but is much less studied than math, logical reasoning, or writing problems. These so-called daily activities can be complex too, by involving multiple-step planning for efficiency, ruling out possibilities in a large search space, using multiple tools in an unconventional manner that even humans find difficult. Namely, solving these "daily activities" requires different kinds of creativity from scientific discovery, art, *etc.*, and have a high potential for AI making people's daily life more enjoyable.

Conclusion We present MacGyver, a novel benchmark focusing on everyday innovation that is carefully collected with quality and diversity control. We evaluate and compare both LLM and human performances, and highlight failure modes of LLMs in proposing physically feasible actions towards a goal. Nonetheless, we find LLM capabilities to be complementary to human capabilities under certain domain-specific settings. We propose two new prompting methods that effectively improve this reasoning ability in LLMs.

G. Future Opportunities

We hope MacGyver dataset opens the door to multiple future directions that will contribute to the broader goal of creating *AI systems that can intelligently and flexibly interact with their surroundings*. For example in this paper, we provide a preliminary attempt to improve the capability of LLMs via two prompting strategies. We encourage future investigation into planning and reasoning strategies to enhance LLMs with physical knowledge and spatial understanding, and to reduce hallucination. To further ameliorate the mistakes made by LLMs in a fictional setting, future work are encouraged to build embodied agents that can interact with physical or simulated worlds and receive feedback from the environment.

Finally, we encourage automatic evaluation methods for this complex reasoning task. For example, using LLMs to extract claims from the candidate solutions, and examine the physical feasibility (or predict the consequences) of proposed actions based on some physical world knowledge.

H. More Information on the MACGYVER Dataset

H.1. Human Verification Process

After generating the challenging scenarios, we involve human verifiers to judge if the final versions of the problems **1)** are solvable (*i.e.,* it is possible to find a reasonable solution using the presented tools), unsolvable, or need more clarification (*i.e.,* the setup is vague or contradictory, which will be discarded), and **2)** for those solvable, whether solving them efficiently requires creative thinking (*i.e.,* using objects to achieve goals they were *not* originally designed for —unconventional usage). Each problem is annotated by three human verifiers from Amazon Mechanical Turk. The detailed verification interface can be found in Appendix I.6. The average inter-annotator agreement (IAA), measured by Cohen's Kappa, are 0.67 and 0.77 for tasks **1)** and **2)**, respectively.

H.2. Collecting Gold Solutions

We provide more details on the final step of our data collection —to pair each problem with a gold answer. For the solvable subset, the answer is a feasible solution written step by step. For the unsolvable subset, the answer is a correct explanation for why the stated goal cannot be achieved.

To save human effort, we start by leveraging the generative strengths of a powerful LLM, *i.e.,* GPT-4. Specifically, we first prompt GPT-4 to generate a solution for each problem in the MACGYVER dataset. Then, human verifiers assess whether the generated solutions are valid. Only if **all three verifiers** agree that a solution is valid, it becomes part of our dataset. Otherwise, we ask human workers to write down a solution (for solvable subset) or a justification (for unsolvable subset).

H.3. Does the data collection pipeline result in progressively challenging problems?

To test whether our data creation pipeline (in Figure 2) is indeed iteratively posing challenge to a previous iteration, we collect GPT-4 answers to iteration 1, 2, and 3 of 200 problems, and run the same human evaluation process.

GPT-4's performance on all three iterations of the same set of problems can be found in Table 7. As the problems get iteratively refined, the ratio of feasible and efficient solutions decrease, and the ratio of infeasible answers increase. This reflects that most potent LLM, GPT-4, indeed finds the problems increasingly challenging.

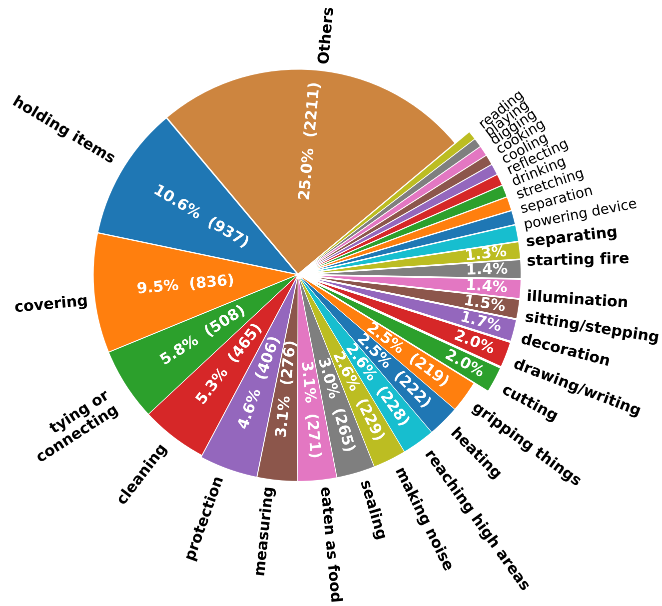


Figure 13. Affordances of the presented tools in our MACGYVER dataset and their frequency (and count). Note that one object may have multiple affordances (*e.g.*, paddle boards can be used for boating, reaching high areas, and exercise).

H.4. Diversity Control and Check

Intuitively, we want to avoid generating multiple problems with familiar goals and constraints. In this section, we summarize our measures to ensure the collected problems are *diverse, comprehensive, and free of repetitive patterns.*

For **diversity control**, We hand-craft more than 50 tags of locations and activities, aiming to ensure that our data collection pipeline delves into a variety of topics. These predefined tags are integrated into the prompt that we used to query GPT-4 for problem curation at Iteration 1. The detailed list of all tags can be found in Table 6.

For **diversity check** After the final iteration, we parse the objects presented as tools among all generated problems. Intuitively, we consider two similar objects with different properties (*e.g., plastic knife* and *metal knife*; *eyeglasses* and *magnifying glass*) to be different. In total, 3,800 unique tools were identified. We compute their frequency and use GPT-4 to analyze their affordances (Appendix Table 8; Figure 13). We found that *holding items* and *covering* are the top two types, followed by *tying or connecting* and *cleaning*. The long tails in both illustrations signify a desirable level of diversity.

Tags used for Diversity Control Before the first iteration, we hand craft more than 50 tags of locations and activities, aiming to ensure that our data collection pipeline delves into a variety of topics. The tags cover diverse range of human activities, from indoor ones such as *home arrangement* and *working in the office*, to outdoor ones such as *hiking*,

Indoors/Household	Neutral	Outdoors
bedroom	at a party	at the beach
closet or storage organization	classroom and university lecture hall	backyard gardening
cooking a complex dish	dog training	beach cleanups, or planning a beach event
dining room	garage	boat trip
fitness workouts	going out for a meal	campsite setting
gym and sports facilities	plants, flowers and garden	city streets and sidewalks
hair styling and care	public speaking	construction work
home improvement	recycling and waste management	desert survival
in a hotel room	school and student activity	exploring a cave
indoors arrangement	school science fair	farm duties
kitchen	science laboratory	forest and jungle
library	swimming	hiking, camping, and traveling
living room	university campus	in the parks
office and work	vehicle maintenance	in the rain
packing things up	weather preparation and response	in the winter
personal grooming and beauty routine		in the zoo
shopping		on the playground
		playing with snow
		playing with water
		rooftop terrace

Table 6. The tags (*i.e.,* locations and activities) used to curate the dataset for diversity control. They can be broadly divided into Indoors/Household, Neutral, and Outdoors.

Solutions	Feasible & eff. (↑)	Feasible & ineff.	Infeasible (↓)	LLM says unsolv. (↓)
Iteration 1	39.1%	36.8%	24.0%	0.1%
Iteration 2	34.7%	32.2%	31.7%	1.4%
Iteration 3	25.4%	37.9%	35.7%	1.0%

Table 7. GPT-4 performance on iteration 1, 2, and 3 of 200 problems. Numbers in each row add up too 100%.

gardening, and *playing with water*. These predefined tags are integrated into the prompt that we used to query GPT-4 for problem curation at Iteration 1. We list all the tags (*i.e.,* locations and activities) used to curate the dataset in Table 6. They are introduced to prompt the LLM for diversity control, and can be broadly divided into Indoors/Household, Neutral, and Outdoors.

Generation in Batch All problems are generated and refined in batches of 15 rather than one by one, as we find out the former results in significantly higher diversity. We then leverage a widely-used sentence transformer (Reimers & Gurevych, 2020) to filter out any newly generated problem that is semantically similar to the existing ones in our database.

Analyzing Tool Affordance We leverage GPT-4 to analyze the affordance of presented tools in the MacGyver dataset. Specifically, we start with a small set of handcrafted affordance as seed. Despite being required to choose only from this fixed list of affordances, GPT-4 does not strictly follow our instruction, and sometimes returns new types that are not included in the seed list. We then gradually expand the list of affordances with newly generated ones.

For eliciting tool affordances, we use the prompt shown in Figure 14.

Commonly-presented tools and their frequencies In total, more than 3,800 different tools appear in our MacGyver dataset. We list in Table 8 16 commonly-presented tools, their featured affordances, and frequency. The number of unique tools and the long tails in distribution signify a desirable level of diversity.

I. Experimental Details

I.1. Collecting Independent Human Responses

We assessed human capability by recruiting participants who are new to this task. To this end, independent solutions were collected from a pool of $N = 252$ UK participants on Prolific. We intentionally used a different platform and target population from those of the human evaluators (*i.e.*, MTurk and US) to minimize any chances of overlap. For a given problem, participants indicated whether they believed the problem is solvable, unsolvable, or required further clarification. If solvable, they provided a step-by-step solution, and otherwise explained why the problem was unsolvable. Overall, we elicited an average of six responses per problem and each participant contribute to up to five different problems.

```
<-- Instruction. -->
You need to write the most common affordances of an item.
    Please choose one or more options from the following
    :
<-- Seed list to expand with. -->
Container/holding items, covering, heating, measuring,
    drawing/writing, cleaning, sitting/stepping, tying or
    connecting, illumination, stretching, starting fire,
    sealing, cutting, separation, reaching high areas,
    powering devices, digging, making noise, flatten,
    cutting, gripping things, reflecting, eaten as food.

<-- Examples. -->
Here are some examples:
 rice: eaten as food
 case: container/holding items, protection, covering
 ruler: measuring, straightening
 box: container/holding items
 pencil: drawing/writing,

<-- Actual Task. -->
Please write the common types of affordances of the
    following tools.

1. {Tool 1}.
 ...
N. {Tool N}.
```

Figure 14. The prompt used to analyze tool affordance. We start with a list of affordances as seed. We gradually expand our list thanks to the fact that GPT-4 does cannot strictly follow our instruction and occasionally generates other affordances not belonging the predefined set.

I.2. Collecting Machine Responses

We collected solutions from seven different LLMs using Nucleus sampling (Holtzman et al., 2020) and return the top one sequence (T=0.7 and p=0.95). In the prompt, we instruct an LLM to either provide a feasible and efficient solution to a problem when it believes the problem is solvable, or otherwise a justification explaining why the given problem is unsolvable. To explore whether different sizes of the same model plays a role in its problem solving ability, we include three variations of Llama2 (*i.e.*, `-7b`, `-13b`, `-70b`), as well as two variants of GPT model family (*i.e.*, `gpt-3.5-turbo`, `gpt-4-0613`).

Additional GPT-4 Responses For a fair comparison with humans, we emulate the same setup in Appendix I.1 by obtaining multiple solutions per problem from a single LLM. Since exhaustive human evaluation is costly, we opted to elicit multiple solutions exclusively from the most capable LLM, GPT-4. Multiple manually-designed instructions are used to prompt GPT-4 in order to reduce repetition among separate sessions of API calls. More details can be found in Appendix I.3.

Tool	Affordance	Freq.
duct tape	sealing; tying or connecting	2.0%
plastic bag	container or holding items; covering	0.7%
flashlight	illumination	0.7%
aluminum foil	covering; heating; sealing	0.6%
hairdryer	heating; drying; making noise	0.5%
ruler	measuring; straightening	0.4%
broom	cleaning; sweeping; reaching high areas	0.4%
spoon	eating; stirring; measuring	0.4%
toothbrush	cleaning; spraying	0.4%
mag. glass	magnifying; starting fire	0.4%
rope	tying or connecting; reaching high areas	0.4%
hammer	flattening; gripping things; making noise	0.3%
yoga mat	stretching; sitting/stepping; covering	0.3%
towel	wetting; covering; absorbing	0.3%
frisbee	playing; throwing	0.3%
toothpick	cleaning; separating	0.3%

Table 8. Examples of most commonly presented tools, their featured affordances, and frequency of these tools in the entire dataset. We randomly pick 16 tools from the top 40 frequent ones in the MacGyver dataset. In total, more than 3,800 different tools appear in our dataset.

I.3. Benchmark Setup

Recruiting MTurk Evaluators We used qualification tasks to recruit 160 qualified annotators on Mechanical Turk. They are paid over 18 USD per hour for all the evaluation and verification tasks.

Collecting Human Solutions on Prolific All participants of human study provide informed consent in accordance with an approved IRB protocol. For a given problem, participants indicated whether they believed the problem is solvable, unsolvable, or required further clarification. If solvable, they provided a step-by-step solution, and otherwise they explained why the problem was unsolvable. A screenshot of the elicitation interface is shown in Figure 22.

Collecting Multiple GPT-4 Responses in Benchmark Recall that in Appendix I.2, we elicit multiple solutions exclusively from the most potent LLM, GPT-4, to emulate the same setup of human study. To align with the varying number of human responses for different problems, we adjusted the quantity of collected GPT-4 answers to match that of human answers. On average, we elicited four GPT-4 solutions per problem through separate API call. To this end, four manually-designed instructions are used to prompt GPT-4 to reduce repetition among separate sessions. For each API call, we still adopt Nucleus sampling and return the top one sequence.

I.4. Analyzing Results

Each machine-generated or human-written answer is annotated by three Mturk workers, with an average IAA of 0.71 as measured by Cohen's Kappa, indicating a substantially strong agreement. Interestingly, we notice that human workers disagree more often when deciding whether a solution is efficient or inefficient. Upon further investigation, we realize this is partially due to the limitation of individual annotator's capability – a person who is unaware of the most efficient solution might label a sub-optimal one as highly efficient. Therefore, for those generated solutions linked to solvable problems, instead of taking the *majority* vote, we take the *worse* labels as the golden label (*e.g.,* taking *'ineff.'* from *['eff.', 'ineff.', 'eff.']*). For all other cases, we still take the majority votes as gold labels. We find such modification leads to a more accurate set of labels.

I.5. The Prompts for Improving LLM's Ability

Figure 15 and Figure 16 list the actual prompts for `Self-Reflection` and `Divergent-Convergent Thinking`.

I.6. Human Task Interfaces

Data Collection and Difficulty Assessment. In practice, we combine the questions of data collection (Section 2) and difficulty assessment (Appendix A) into one single task. The detailed human annotation interface, including the instructions, examples, and the actual task and be found in Figure 17 to Figure 21.

Human Study A screenshot of the interface to elicit independent human responses is shown in Figure 22. For a given problem, participants indicate whether they believe the problem is solvable, unsolvable, or required further clarification. If solvable, they provide a step-by-step solution, and otherwise they explain why the problem was unsolvable.

Benchmark Evaluation The screenshots of our human evaluation interface for the benchmark experiment can be found in Figure 23 and 24.

```
<-- Round 1: -->
User: {Problem Statement}
If the problem is solvable, provide a concise solution. Use step1, step2, etc, and mention the tools to achieve each
    step. Use as few steps as possible and the answer should ideally be less than 100 words.

If you cannot find a feasible solution, just say that it is not possible and give a very short justification.

Assistant: {Answer}

<-- Round 2: -->
User: Now, please verify if each step is physically feasible and afforded. After that, modify the solution if needed.
Use the following format:
Step 1: ...
Step 2: ...
...
Conclusion 1: Whether the problem is indeed solvable given all the constraints
Conclusion 2: (If still solvable) No modification needed/Modification needed.


Modified solution:
Assistant: {Response and Updated solution}
<-- Repeat until no modification is needed.-->
```

Figure 15. Prompt used for the `step-by-step verify` strategy.

```
User: {Problem Statement}
Give a feasible solution very concisely. Note that some tools are not useful, so please analyze the affordance of each
    presented object, and rule out unnecessary ones first.


Use the following format:
1. List the affordance of presented items and whether they are useful
2. Summary: list useful tools
3. If the problem is solvable under all these constraints, write the solution. Use step1, step2, etc, and mention the
    tools to achieve each step. Use as few steps as possible and the answer should ideally be less than 100 words.

If you cannot find a feasible solution, just say that it is not possible and give a very short justification.

Assistant: {Analysis of the affordance and the main answer}
```

Figure 16. Prompt used for the `divergent-convergent thinking` strategy.

Instructions (Verify Problems and Solutions)

Thanks for participating in this HIT! Please read the instructions carefully.

In this task, you will be presented with a problem and a solution. Your task is to 1) judge by yourself if the problems are solvable (i.e., you can find a **reasonable and safe** solution to the problem using the presented tools), and **2) judge if solving the problem requires using tools in an unconventional way**, and 3) verify if a proposed solution is **feasible, efficient** or not.

- For those you select as ***unsolvable problems***, you will be asked to provide an optional short explanation to justify your choice. Next, you will read an answer verify its correctness.
- For those you select as ***solvable problems***, you will be asked to decide if solving the problem efficiently requires using tools in an unconventional way. After that, you will read a potential solution and judge if the presented solution is feasible, efficient, and uses tools unconventionaly.

Here are the definitions and examples of new concepts:

Solvable and Unsolvable Problems:

- A problem is **Solvable** if and only if you can achieve the stated goal 1) using only the presented tools, and 2) in a safe and reasonable way.
- For example, the problem below is **unsolvable** because Car engine is the only available tool that provides source of heat, but it is not safe to use car engines to cook food.

  At a food truck park, your takeout food is getting cold and you want to reheat it. Tools available: car engine, foil paper that is very thin and tears easily, a rock, a can of soda, a cloth bag. How do you reheat the food taking into account the condition of available tools?
- For example, the below problem is **ambiguous** because It is not clear why the keys on a keyboard stopped working, thus unclear how to solve the problem.

  A few keys on your keyboard have stopped working, and you don't have a replacement. You have a pencil sharpener, a box of push pins, a booklet of envelopes, a rubber band, a paperclip, a sticky note pad, and a glue stick, but the sticky notes have lost their stickiness and the paperclip is too stiff. How can you temporarily fix the problem?

Feasible/Infeasible/Efficient/Inefficient Solutions:

- **Feasible:** A solution is feasible if it 1) proposes a valid approach to solve the stated problem, and 2) each step is achievable using the presented tools.
- **Infeasible:** A solution is infeasible if 1) the proposed approach doesn't solve the stated problem, or 2) at least one step is NOT achievable.
- **Efficient**: A solution is efficient if 1) it has no redundant or unnecessary steps, and 2) it is unlikely that the problem can be solved using a fewer number of steps.

Conventional and Unconventional Usage of Tool:

- **Conventional Usage of Tools** refers to the traditional and established ways to use a tool or item. Items are specifically designed to fulfill such goals effectively.

  For example, using a broom to sweep the floor is a conventional usage.
- **Unconventional Usage of Tools**: refers to the unusual ways to utilize a tool or item. Items are NOT orginially designed to fulfill such goals.

  For example, using the same broom's broomstick to reach a high place is an unconventional usage.

Figure 17. Human Annotation Interface for Data Collection and Difficulty Assessment, Page 1.

Example with Explanation (click to expand/collapse)

If you haven't, please carefully read the following problem-solution pairs below.

1. Example of Efficient/Inefficient/Infeasible Solution

Example Problem: On a desk placed against the wall, you are given a candle, a sturdy paper box without a top (roughly the size of your hand), plenty of thumbtacks, and a box of matches. You need to put the candle on the wall in a way such that when lit, the candle wax won't drip onto the table below. How would you process with only these items?

Solution 1:
Step1: Use the matches to light the candle.
Step2: Melt the candle's wax and use the melted wax as an adhesive to stick the candle on to the wall.
Step3: Hold the candle by hand until the wax cool down.
Step4: Place the paper box on the desk, right below the candle to catch the wax.

Answer 1: **Feasible but inefficient**. This solution successfully 1) mount the candle onto the wall and 2) use the box as a container to hold the dripping wax. However, it is not the most efficient solution (see Solution 2).

Solution 2:
Step1: Pin the paper box onto the wall using the thumbtacks.
Step2: Put the candle on the paper box. Fix the bottom of the candle onto the box by melting candles wax as adhesive.

Answer 2: **Feasible and efficient**. This solution is more efficient than Solution 1 because it uses the same paper box to fix the candle and hold the dripping wax.

Solution 3:
Step1: Use the thumbtacks to attach the candle directly to the wall.
Step2: Place the box on the table right below the candle. The base of the box should be facing you.

Answer 3: **Not feasible**. There are two problems with this solution. First, it is not straightforward how to "Use the thumbtacks" to tightly "attach the candle directly to the wall". Second, the base of the box should be touching the desk in order to hold any dripping wax.

Figure 18. Human Annotation Interface for Data Collection and Difficulty Assessment, Page 2.

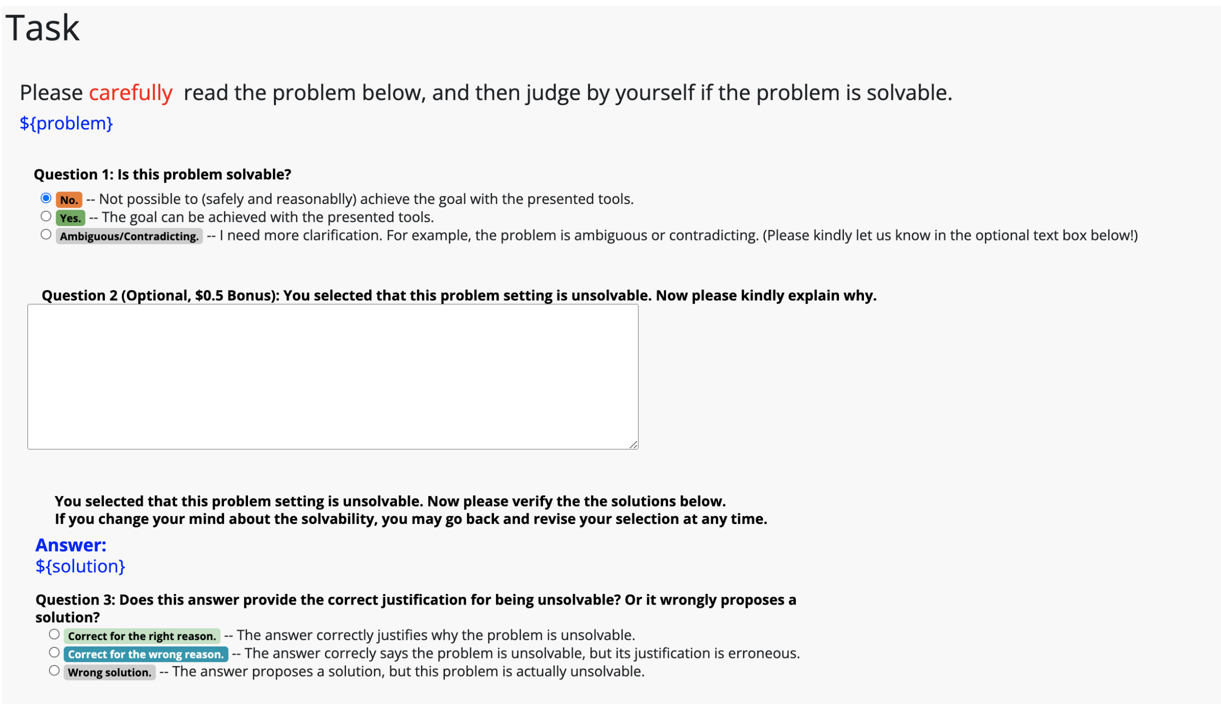
2. Example of Unconventional Usage of Tools

Problem 1: On a table leaning against the wall, you are given a candle, a box of thumbtacks, and a box of matches. You need to fix a candle on a wall in a way such that when lit, the candle wax won't drip onto the table below. How would you proceed?
Solution 1:
Step 1: Take out the thumbtacks from the box.
Step 2: Use the thumbtacks to pin the box that's used to hold the thumbtacks onto the wall.
Step 3: Put the candle on the paper box. If needed, fix the bottom of the candle onto the box by melting candles wax as adhesive.
Answer 1: The solution thinks out-of-the-box, by emptying the box which served as a container of the thumbtacks and use it to hold the candle. Therefore, in the above example, the box (which served as a container of the thumbtacks) is being used **unconventionally**.

Problem 2: You need to roll out dough evenly but don't have a rolling pin. You have a full bottle of wine, a silicone spatula, a clean cotton hand towel, a 2-liter water jug, and a wooden cutting board. How could you proceed?
Solution 2:
You can use a full bottle of wine as a makeshift rolling pin.
Answer 2: We know that glass bottles of wines are NOT originally intended to roll a dough. Therefore, in the above example, it is being used **unconventionally**.

Problem 3: You have a balloon, a bottle of superglue, and a corkboard. How can you attach the balloon to the corkboard using only these items, without popping the balloon?
Solution 3:
Step 1: Apply a small amount of superglue to the corkboard.
Step 2: Gently press the balloon onto the glued spot.
Step 3: Hold the balloon in place briefly to allow the glue to bond.
Answer 3: We know that glue is invented to adhere or join things together. Therefore, in the above example, the superglue is being used **conventionally**.

Figure 19. Human Annotation Interface for Data Collection and Difficulty Assessment, Page 3.



Task

Please carefully read the problem below, and then judge by yourself if the problem is solvable.
${problem}

Question 1: Is this problem solvable?

- No. -- Not possible to (safely and reasonablly) achieve the goal with the presented tools.
- Yes. -- The goal can be achieved with the presented tools.
- Ambiguous/Contradicting. -- I need more clarification. For example, the problem is ambiguous or contradicting. (Please kindly let us know in the optional text box below!)

Question 2 (Optional, $0.5 Bonus): You selected that this problem setting is unsolvable. Now please kindly explain why.

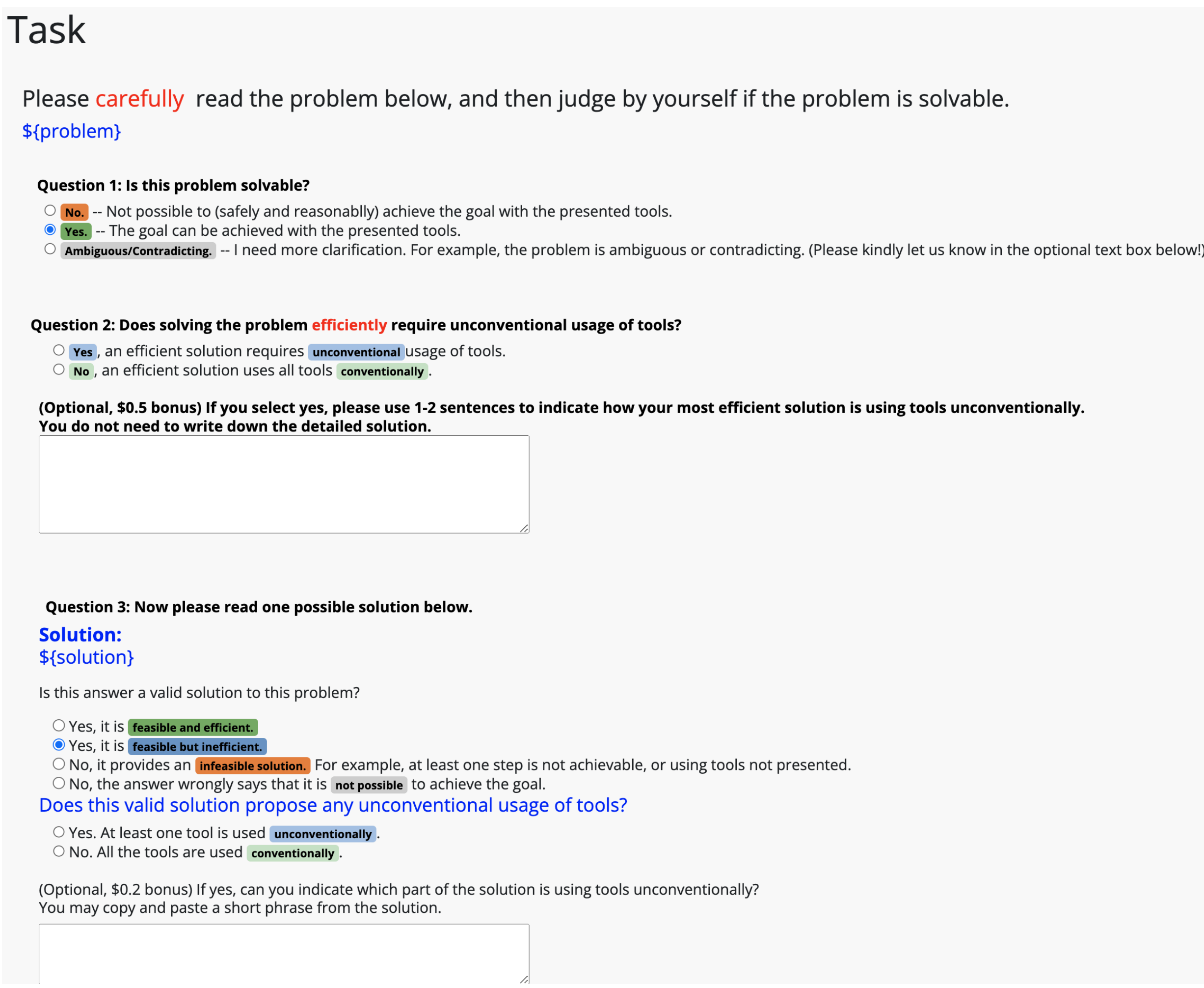
You selected that this problem setting is unsolvable. Now please verify the the solutions below.
If you change your mind about the solvability, you may go back and revise your selection at any time.

Answer:
${solution}

Question 3: Does this answer provide the correct justification for being unsolvable? Or it wrongly proposes a solution?

- Correct for the right reason. -- The answer correctly justifies why the problem is unsolvable.
- Correct for the wrong reason. -- The answer correcly says the problem is unsolvable, but its justification is erroneous.
- Wrong solution. -- The answer proposes a solution, but this problem is actually unsolvable.

Figure 20. Human Annotation Interface for Data Collection and Difficulty Assessment, Page 4.



Task

Please carefully read the problem below, and then judge by yourself if the problem is solvable.
${problem}

Question 1: Is this problem solvable?

- ○ No. -- Not possible to (safely and reasonablly) achieve the goal with the presented tools.
- ◉ Yes. -- The goal can be achieved with the presented tools.
- ○ Ambiguous/Contradicting. -- I need more clarification. For example, the problem is ambiguous or contradicting. (Please kindly let us know in the optional text box below!)

Question 2: Does solving the problem efficiently require unconventional usage of tools?

- ○ Yes, an efficient solution requires unconventional usage of tools.
- ○ No, an efficient solution uses all tools conventionally.

**(Optional, $0.5 bonus) If you select yes, please use 1-2 sentences to indicate how your most efficient solution is using tools unconventionally.
You do not need to write down the detailed solution.**

Question 3: Now please read one possible solution below.

Solution:
${solution}

Is this answer a valid solution to this problem?

- ○ Yes, it is feasible and efficient.
- ◉ Yes, it is feasible but inefficient.
- ○ No, it provides an infeasible solution. For example, at least one step is not achievable, or using tools not presented.
- ○ No, the answer wrongly says that it is not possible to achieve the goal.

Does this valid solution propose any unconventional usage of tools?

- ○ Yes. At least one tool is used unconventionally.
- ○ No. All the tools are used conventionally.

(Optional, $0.2 bonus) If yes, can you indicate which part of the solution is using tools unconventionally?
You may copy and paste a short phrase from the solution.

Figure 21. Human Annotation Interface for Data Collection and Difficulty Assessment, Page 5.

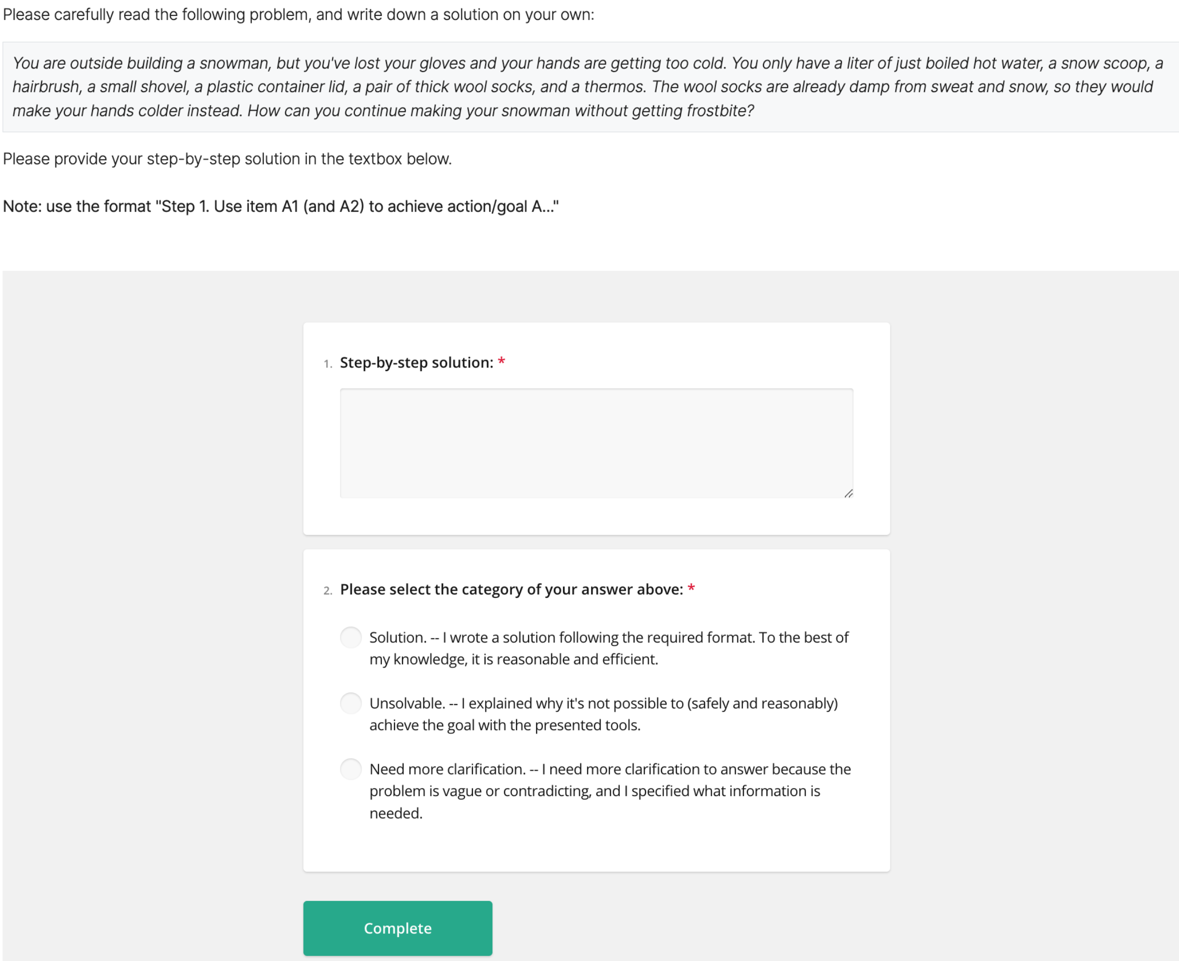


Figure 22. Human Study Interface to Collect Independent Human Responses.

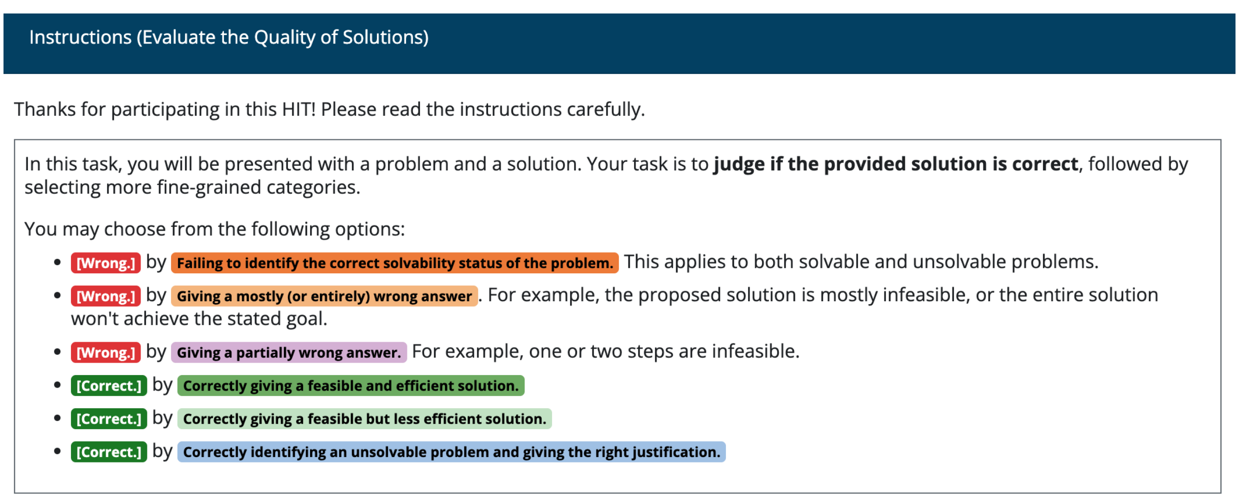


Figure 23. Human Evaluation Interface for Benchmarking, Page 1.

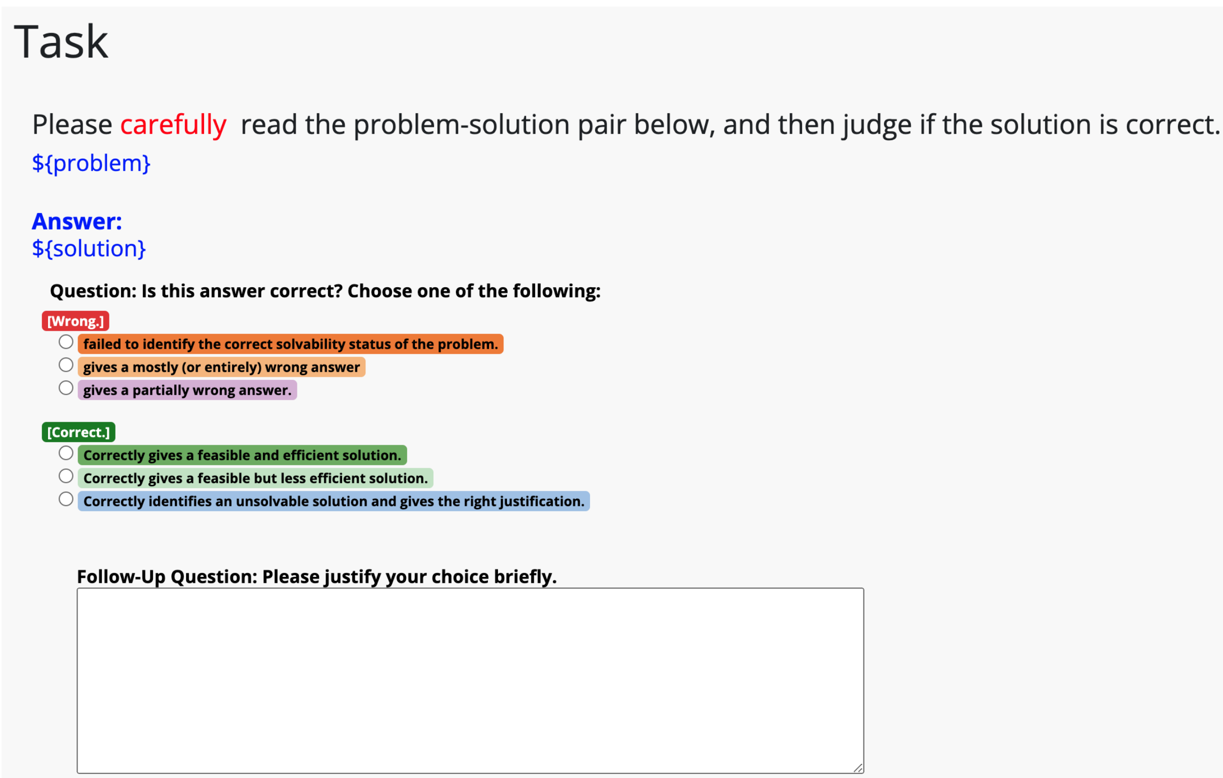


Figure 24. Human Evaluation Interface for Benchmarking, Page 2.