

True Detective: A Deep Abductive Reasoning Benchmark Undoable for GPT-3 and Challenging for GPT-4

Maksym Del and Mark Fishel
Institute of Computer Science
University of Tartu, Estonia
{maksym,mark}@tartunlp.ai

Abstract

Large language models (LLMs) have demonstrated solid zero-shot reasoning capabilities, which is reflected in their performance on the current test tasks. This calls for a more challenging benchmark requiring highly advanced reasoning ability to be solved. In this paper, we introduce such a benchmark, consisting of 191 long-form (1200 words on average) mystery narratives constructed as detective puzzles. Puzzles are sourced from the "5 Minute Mystery" platform and include a multiple-choice question for evaluation. Only 47% of humans solve a puzzle successfully on average, while the best human solvers achieve over 80% success rate. We show that GPT-3 models barely outperform random on this benchmark (with 28% accuracy) while state-of-the-art GPT-4 solves only 38% of puzzles. This indicates that there is still a significant gap in the deep reasoning abilities of LLMs and humans and highlights the need for further research in this area. Our work introduces a challenging benchmark for future studies on reasoning in language models and contributes to a better understanding of the limits of LLMs' abilities.¹

1 Introduction

Large language models (LLMs) have gained significant attention in recent years due to their impressive performance on a wide range of natural language processing tasks, including reasoning tasks (Srivastava et al., 2022; Wei et al., 2022). This calls for new, genuinely challenging benchmarks requiring LLMs to possess truly advanced reasoning capabilities to be solved.

Abductive reasoning is a type of inference aiming at finding the minimal and most justified explanation for the set of phenomena or observations. Previous benchmarks on this topic, such as Mostafazadeh et al. (2016), consisted of short

and straightforward common-sense observations and were solved by GPT models (Radford and Narasimhan, 2018). However, the canonical example of abductive reasoning, a demanding process of a detective finding the best solution to a complex crime based on the clues and observations, was not explored as a foundation for the LLM benchmark in the literature.

Motivated by the need for a new reasoning benchmark and inspired by the complexities and particularities of a detective enterprise, we present a novel abductive reasoning benchmark consisting of 191 detective puzzles/mysteries. Mysteries are sourced from the "5 Minute Mystery" platform, where professional and aspiring authors wrote them. A puzzle is structured as a >1000 words story with 4-5 answer options. Over the last 15 years, puzzles were attempted by humans around 2000 times each with an average solve rate of 47% (only the first try for each human for each puzzle counts). However, top human solvers (top 10) achieve a success rate of over 80% solving more than 154 of 192 puzzles correctly.

Moreover, additional modifications such as chain-of-thought (CoT) prompting Wei et al. (2022); Kojima et al. (2022) that are meant to invoke emergent reasoning abilities in LLMs do not help for GPT-3.

In this study, we also assess the performance of the current state-of-the-art GPT-3 and GPT-4 models on our newly proposed dataset. We show that these models, even equipped with the Chain of Thought prompts (Wei et al., 2022; Kojima et al., 2022), are getting an accuracy rate of only 28%, barely better than random guessing (GPT-3.5), or scoring 38% (GPT-4), which is halfway between random guessing and average human baseline, and far behind top human solvers with their 80% solve rate. These results reveal a significant gap in the reasoning abilities of GPT models and humans.

In our ablation study, we also supply models

¹<https://github.com/TartuNLP/true-detective>

with golden CoTs. Golden CoTs are narratives that represent the reasoning behind the correct answer for each story (written by the mystery authors). When we attach golden CoTs to the input prompt, the best-performing GPT-3.5 model only achieves a solve rate of 63%. This indicates LLMs' difficulty making even trivial inferences from the complex long-form story. GPT-4 models, however, get as good as the best human solvers when presented with our chain of thoughts (even though humans do not have access to the golden CoTs).

Our contributions in this paper are twofold: (1) a new challenging benchmark for evaluating LLMs for advanced abductive reasoning; (2) a showcase of GPT-3.5 and GPT-4 models failing to perform reasonably.

2 Related Work

Mostafazadeh et al. (2016) introduced the ROCStories benchmark: narrative cloze test, which requires choosing the correct ending of the four-sentence story. Bhagavatula et al. (2020) expand on this dataset, requiring finding plausible explanations for narrative gaps instead of focusing on the sequence of events. Our benchmark contains stories of around 70 sentences that require solving the detective mystery (as opposed to simply figuring out commonsense story continuation), which is a much harder inference.

Natural language inference (NLI) is another related domain, but NLI tasks usually include much simpler and smaller inferences (Bowman et al., 2015; Williams et al., 2018). Zellers et al. (2018) introduced the SWAG dataset that offers a large-scale natural language inference challenge where grounded knowledge is required to make an inference. This shares some commonality with our dataset, as some mysteries might require a share of grounded knowledge about the real world. Unlike Zellers et al. (2018), we only offer a test set, but our stories are broader and more involved. On the other hand, Grimm and Cimiano (2021) introduced a question-answering benchmark that requires deeper text understanding based on the football match commentaries. Their questions range from counting the number of goals to identifying the game-winner. While answers to many of these questions are not explicitly provided in the football commentary, our mysteries require solving the whole case specifically designed to be challenging even for humans.

Lastly, Wei et al. (2022) find that while eliciting "Chain of Thought" reasoning helps with stronger models, it can hurt when solving harder tasks with smaller models. We observe this behavior when comparing GPT-3.5 and GPT-4 on our benchmark.

3 Benchmark

3.1 5 Minute Mystery Platform

The data for this AI research was obtained from the "5 Minute Mystery"² online platform. This website is an online platform that has functioned for over ten years and allows users to submit and solve mysteries of varying difficulty (see Appendix A for an example mystery).

Based on the website author guidelines, the mysteries on the website collection are intended for readers at the sixth to eighth-grade reading level and have a recommended length of around 1200 words. To facilitate comprehension and challenge the reader, each mystery includes around four suspects and one guilty suspect. Of the 191 mysteries, the overwhelming majority ask the reader to identify the guilty suspect, with only occasional ones asking for the geographic location or the missing person. The aim is for the reader to demonstrate their abductive reasoning abilities by solving the mystery and identifying the correct solution (e.g., the murderer). Typically, one character in the story is faced with the key puzzle, and at the end of the mystery, they exclaim something like: "I figured out who is guilty!" At this point, the reader must choose the correct answer from a list of options.

In addition, mystery writers provided an explanation for the answer: a full solution (golden CoT) that elicits reasoning that leads to the correct answer. The reasoning is presented on behalf of one of the story characters (the one who says, "I know who did it" at the end of the story).

The website also has a unique scoring system that rewards users for correctly solving mysteries, encouraging participation and engagement. In addition to providing entertainment, the website can also be used in an educational setting to help students develop their comprehension and critical thinking skills.

3.2 Benchmark Dataset

The mysteries in this study were obtained from the "5 Minute Mystery" (5MM) platform. We have included links to the original mysteries and to the

²<https://www.5minutemystery.com/>

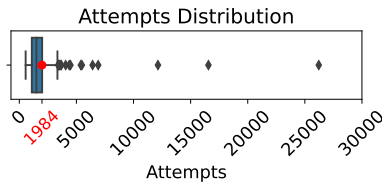


Figure 1: Distribution of the number of attempts for each mystery. The red dot indicates that almost 2000 people attempted mysteries on average. This suggests that our dataset provides a robust estimate of human performance and is representative of human performance on the mysteries.

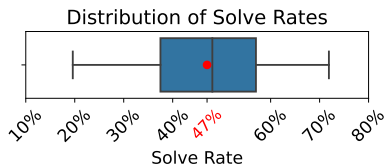


Figure 2: Average human solve rate for each mystery in the dataset. The performance for most puzzles is around 40-60%. The red dot indicates the average solve rate. This figure reveals that the majority of puzzles are challenging for human solvers, providing a good benchmark for evaluating the performance of AI models on these types of tasks.

author pages in the study, and we want to emphasize that all copyrights remain with the original authors and the 5MM team. See the authors list in the Appendix B section.

Dataset size and the number of answers. The dataset used in this study consists of 191 puzzles, including 160 puzzles with four answer options, 30 puzzles with five answer options, and one puzzle with three answer options.

Attempts. The "5 Minute Mystery" platform has been in operation for approximately 14 years and has attracted thousands of users, with over 20,000 registered by 2013. These users have made numerous attempts at each mystery, but only their first attempt is counted towards the platform's statistics. As shown in Figure 1, the average number of attempts per mystery is 1984, with only a few puzzles being significantly more or less popular.

With such a large sample size, the resulting human performance estimate is highly robust and reliable as a benchmark.

Human Solve Rate. In the 5MM platform, human solvers have achieved moderate success. The average solve rate is 47%, significantly higher than

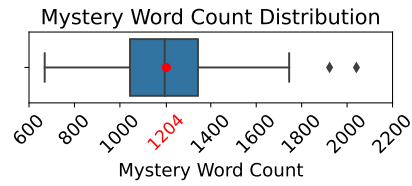


Figure 3: Number of words in each mystery in the dataset. Mysteries range from 600 to around 2000 words with most of them being around 1204 words (red dot). This suggests that not only does the task require drawing highly nontrivial conclusions from the text but also doing so over relatively large texts.

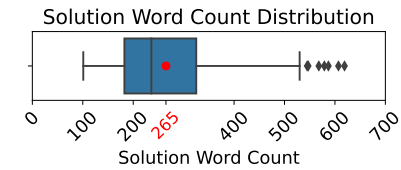


Figure 4: Number of words in the solution explanations for each mystery. The red dot indicates the average number of words per explanation. This figure reveals that the average solution length is 265 words, and the longest solutions are around 600 words. Solutions (golden CoTs) are useful for a setup testing the ability of LLMs to do a trivial final answer inference over the given CoT.

random guessing (around 24%), indicating that the tasks are challenging even for humans. The top ten human solvers have an average solve rate of 80-90%, per platform statistics. Figure 2 shows that most mysteries are solved between 40% and 60% of the time, with some being solved up to 70% of the time and others close to random guessing. While the mysteries were designed to vary in difficulty, it is possible that the best explanation provided by humans may not always align with the author's intended solution for the hardest ones. However, we continue to include these mysteries in our dataset to investigate whether language models can better infer the author's intent in these cases.

Mystery word count. Figure 3 shows the distribution of the number of words in each mystery in the dataset. On average, mysteries have 1204 words, with some being as long as 2000 words. This suggests that the puzzles used in the study not only require advanced reasoning skills to solve but also require finding relevant clues from a relatively long body of text that can incriminate or exonerate suspects. This further complicates the task.

Golden CoTs. Each mystery in the dataset includes a full-text solution that provides an expla-

nation of how one of the story characters came up with the correct answers. The average length of these solutions is 265 words, as shown in Figure 4. The solution lengths do not vary significantly, with the longest solution being around 600 words.

These solutions can be considered as ground-truth Chains-of-Thought (cite paper here), which provide insight into the author’s reasoning for each puzzle. This information is valuable for a few reasons. First, it can be used as part of few-shot learning examples (again, cite). Second, as we demonstrate in Section 4, we can use these Chains-of-Thought to simplify the abductive reasoning task and evaluate whether language models can perform inference when the solution is strongly hinted at.

4 Evaluation

4.1 Models

The models used in this study are the InstructGPT-3.5 models *GPT-3.5 (FeedME)*, *GPT-3.5 (PPO)* (OpenAI, 2022), and *GPT-4* (OpenAI, 2023). They are causal language models based on the Transformer architecture (Vaswani et al., 2017) featuring supposedly around 175B parameters for GPT-3.5s.

GPT-3.5 (FeedME): a model was trained using the FeedME method, a supervised fine-tuning method based on human-written instructions and model samples (Ouyang et al., 2022; OpenAI, 2022).

GPT-3.5 (PPO): is a more performant update over *GPT-3.5 (FeedME)* model. Apart from instruction tuning, it was also calibrated with RLHF, a reinforcement learning method that uses reward models trained from human comparisons (Stiennon et al., 2020; OpenAI, 2022).

GPT-4: state-of-the-art commercial model from OpenAI. Achieves human parity on multiple extremely challenging tasks (OpenAI, 2023).

4.2 Methods

In this study, we tested GPTs in a zero-shot manner in three scenarios. This subsection outlines them.

Vanilla: This method involves the task description, mystery body, and an immediate request for the final answer (Brown et al., 2020).

CoT: This method asks LLMs to generate a Chain-of-Thought first (Wei et al., 2022; Kojima et al., 2022) and only then requests the final answer. Chain-of-thought, if reasonable, allows the

model to approach complex problems gradually and unlocks strong reasoning abilities at a particular model scale (Wei et al., 2022).

Golden CoT: This method involves generating answers to instruction-based questions by using a set of ground-truth Chain-of-Though solutions included as part of the prompt. This significantly simplifies the task for the model as it does not need to come up with CoT, so we can test how much of the performance depends on the CoT and how much on the final abductive reasoning step.

4.3 Prompt Templates

Figure 5 shows the task instruction that we give to the InstructGPT models at the beginning of the prompt.

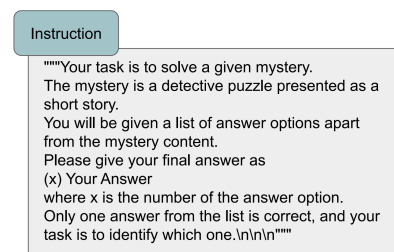


Figure 5: Task instruction that we use as a prompt prefix.

Then we always add the mystery name, list of suspects, and mystery content (body) to the prompt.

When we want to invoke Chain-of-Though reasoning, we also append the following:

```
Full answer:
Let's think step by step.
```

When we want to provide a golden Chain-of-thought, we append the following prompt:

```
Solution:
{solution}
```

Finally, we always ask for the final answer with

```
Final answer:
```

4.4 Results and Discussion

The evaluation results shown in Table 1 indicate that the performance of both *davinci* models under both *Vanilla* and *CoT* prompting scenarios is close to random. In our analysis, we also found that there is no correlation between the length of the mystery or human solve rate and the GPT’s correctness.

Our Golden CoT ablation study (Table 1) demonstrates that even with relevant explanatory CoT, GPT-3.5s can only solve 63% of puzzles correctly, suggesting that difficulty lies not only in generating

Method	Solve rate
Random guess	0.24
Human average	0.47
Human top	0.8-0.9
Vanilla	
GPT-3.5 (FeedME)	0.28
GPT-3.5 (PPO)	0.26
GPT-4	0.27
CoT	
GPT-3.5 (FeedME)	0.26
GPT-3.5 (PPO)	0.29
GPT-4	0.38
Golden CoT*	
GPT-3.5 (FeedME)	0.46
GPT-3.5 (PPO)	0.63
GPT-4	0.83

Table 1: Performance of GPT-3.5 (FeedME), GPT-3.5 (PPO), and GPT-4 under different prompting scenarios against the human baseline. Both vanilla task formation (Instruction and immediate answer request) and "step-by-step" chain-of-thought approaches perform almost equivalent to random guess. Even in unfair comparison, GPTs cannot match/outperform top human solvers when provided with golden chains of thought.

the correct theory for the crime but also in making final inferences when all information is available. On the other hand, GPT-4 does not help such a problem with 83%.

CoT performance of GPT-3.5 models show small to no gains in performance compared to Vanilla. As indicated in Wei et al. (2022), a similar decrement (between GPT-3 and smaller models) was observed in models that weren't sufficiently powerful for the task suggesting that the GPT-3.5 models might also not be strong enough to generate CoT chains that would benefit the task. On the other hand, CoT GPT-4 performs better, although still underachieving compared to the average human solve rate.

The complexity of the long-form multi-character narrative and the level of reasoning required to solve the detective puzzle makes our benchmark especially difficult and sets it apart.

Finally, we explore the complexity of the cases that GPT-4 (CoT) found easier or harder to manage. In our study, we did not observe a direct correlation between the length of a mystery and the level of difficulty it presented. However, when considering the level of concurrence between human decisions and those made by GPT-4, Figure 6 demonstrates a considerable degree of agreement. Specifically, the cases perceived as challenging or straightforward by the GPT-4 were often viewed similarly by human subjects.

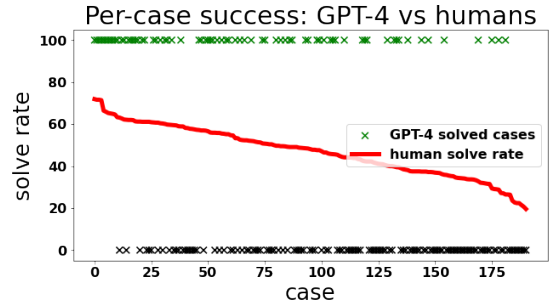


Figure 6: Red line indicates case difficulty for humans, green points indicate cases where GPT-4 (CoT) solved the case successfully, and black points are for cases where GPT-4 failed. Black points are crowded on the right and green points are crowded on the left which correlates with hard and easy cases (as per humans) respectively. Therefore, GPT-4 and humans find similar cases easy/difficult.

5 Conclusion

We presented a new benchmark in the form of detective puzzles to evaluate the abductive reasoning capabilities of Large Language Models. Results from state-of-the-art GPT-3.5 models across three prompting strategies showed poor performance close to random. GPT-4 managed to show comparably solid performance (when prompted with CoT), but even this model is behind the average human solve rate on our benchmark. When provided with golden CoTs, which significantly simplifies the task, GPT-4 shows good performance, while GPT-3 is still unable to do a final inference well enough. Overall, our benchmark offers insights into LLMs' limitations and provides a difficult challenge for future research on abductive reasoning in large LMs.

6 Limitations

Our evaluation focused solely on the performance of leading-edge GPT models details and weights of which are not publicly available. However, there is potential value in extending this study to incorporate other models like PaLM (Chowdhery et al.) or LLaMA (Touvron et al.), which we have earmarked for future research.

Also, as the performance for average humans is only 47% it is possible that some mysteries are ill-defined or unreasonably complicated. Among the top 10 human solvers, the solve rate is also only around 80-90%, and GPT-4 only solves 83% of tasks when provided with ground truth CoTs which drastically simplifies the task.

References

- Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. [Abductive commonsense reasoning](#). In *International Conference on Learning Representations*.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pilla, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#).
- Frank Grimm and Philipp Cimiano. 2021. [BiQuAD: Towards QA based on deeper text understanding](#). In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 105–115, Online. Association for Computational Linguistics.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#).
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. [A corpus and cloze evaluation for deeper understanding of commonsense stories](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.
- OpenAI. 2022. [Model index for researchers](https://beta.openai.com/docs/model-index-for-researchers). <https://beta.openai.com/docs/model-index-for-researchers>.
- OpenAI. 2023. [Gpt-4 technical report](#).
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. [Training language models to follow instructions with human feedback](#). *ArXiv*, abs/2203.02155.
- Alec Radford and Karthik Narasimhan. 2018. [Improving language understanding by generative pre-training](#).
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabasum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito,

Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kočoň, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chifullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkienn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Śwędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimeo Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Moshdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans,

Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Mishnerghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.](#)

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. [Learning to summarize with human feedback.](#) In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard

Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#).

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. [SWAG: A large-scale adversarial dataset for grounded commonsense inference](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

A Example: The Easter Egg Mystery

This appendix provides the most attempted mystery under 700 words as an example. Copyright belongs to the mystery author.

Metadata

- Mystery Name: The Easter Egg Mystery
- Author: Tom Fowler³
- Solve Rate: 60.8%
- Attempts: 1871
- Answer options: (a) Anna; (b) Cole; (c) Justin; (d) Lizzie; (e) Rachel.

Mystery Body Karen Sheldon had loved Easter egg hunts ever since she was a little girl. That is why she eagerly volunteered to assist with this year’s Hunt for the children at her church.

This year, the Children’s Day Out mothers decided to do something different. Because there were so many children of all ages in the congregation, they split the hunt up into age groups. Karen’s job was to oversee several of the 6-10 year olds.

³<https://www.5minutemystery.com/author/tfowler>

Within her group were five children she knew well. They were Rachel Smithson, whose mother Karla had volunteered to help a very grateful Karen, Justin Bates, a classmate of Rachel’s, Karen’s daughter Lizzie, Lizzie’s best friend Anna Laughlin and Cole Bryant, who was also the Sheldon’s next door neighbor.

The Easter egg hunt was on Saturday morning, the day before Easter Sunday. It was held in the large field in back of the church. Karen and Karla were grateful that today was sunny and warm although it was a bit windy. Karen was excited as the children prepared for the hunt, which was to begin at 10:00 am and last for one hour. Just before the start whistle blew, Karen told the children, “I have placed a golden Easter egg in our hunting area. There is an extra bag of candy for the child who finds it.” Only Karla and she knew that the golden egg was placed in back of the largest tree in the field, an old oak in the far corner to the left of where she and the children now stood and an area dedicated to the 6-10 year old age group.

During the hunt, Karen and Karla visited while they watched the egg hunt. During the hunt, Karen noticed that Cole stayed focused on the evergreen shrubbery in the middle of the field, finding several eggs there, much to his delight.

Karen was amused when Rachel ran to her mother and told her, “I have found a lot of eggs. I’m heading back to the rock pile. I bet I will find the golden egg there!” The rock pile was to the right of the evergreen shrubbery.

In the middle of the hunt, Karen excused herself to go inside the church to get a drink of water and sit for a few minutes. When she returned, Karla told her, “I had to run over and warn Lizzie to be careful of the dead branches on the big oak tree. One of them fell last week, hitting one of the older kids.”

As the hunt began to wind down, Karla walked out to speak with a very agitated Anna. After returning to Karen, she told her, “Anna is upset because she has found only a few eggs. I told her to keep looking; there are still a few minutes to go.” Karen noticed that Anna stayed close to Karla for the remainder of the hunt.

As the whistle blew to end the hunt, Karen walked to the center of the field to wave Justin back in. He was in the far right corner of the field, where he had been for the entire hunt. There was a sand pit in that area and Justin found several eggs

there.

As the kids headed back to the start area, Karen once again excused herself to go inside. The wind had blown a speck of dust in her eye when waving Justin down and it was very painful. When she returned from rinsing her eyes, Karla and the five children were smiling at her. She asked, "What's up?"

Karla answered, "One of our kids found the golden egg. We want you to guess which one."

Karen smiled in return, saying, "So that's it!" Thinking for a moment, she said, "I only have one question. When I was inside the first time, did any of the children move from one side of the field to another?"

Karla answered, "No."

Karen tousled Justin's hair and said, "Good. Then I know who has the golden egg!"

A.1 Golden CoT and Answer

Golden CoT. "Good naturedly, Karla exclaimed, "How do you know?" Smiling at Anna, she answered, "It's not too hard to figure out. Let me explain." The eyes of all of the children and Karla were upon her as she continued, "I placed the golden egg behind the big oak tree." Smiling next at Cole, she said, "Cole spent the entire hour in the shrubbery, in the middle of the field, far away from the oak tree." She patted Rachel's shoulder and said; "Rachel did all of her hunting in the rock pile, even farther away from the oak tree." Looking back at Anna, Karen said, "I know you don't have the golden egg, sweetie. You were upset that you had so few eggs with only a few minutes left in the hunt and stayed close to Karla until the whistle blew." Patting her hand, she added, "I'm sure you will do better next year." Turning to Justin, Karen said, "You were farther away from the oak than anyone. You spent the whole hour far out in the sand pit. I even had to come get you because you could not hear the whistle." All eyes turned toward Lizzie. Her mother said, "So, you must have the egg. Karla told me she had to warn you of the dead branches on the oak. You were the only one near it." Pausing, she added, "I hope everyone believes that I did not tell you where I put that egg! Karla jumped in, "Of course we do not think that!" All of the kids echoed their support. Lizzie broke the silence. She said, "I didn't know about the egg until Mother told everyone else before the hunt." Walking over to her side, Lizzie looked at Anna and offered her the golden

egg, saying, "I would like for you to have this." Tearfully, Anna thanked her friend, saying, "This is the best Easter egg hunt ever!" Karen was so proud of Lizzie that she heartily agreed with Anna."

Answer: (d) Lizzie

B Mystery Authors Acknowledgment

We thank all the 5MM mystery writers: Moe Zilla, Tom Fowler, William Shepard, Laird Long, Robbie Cutler, Barney Parmington, Stefanina Hill, Steve Shrott, Nick Andreychuk, Nicholas LeVack, Ernest Capraro, Andrea Hein, Doug Fellin, Tammy-Lee Miller, Meghan Ford, Brad Marsh, Susanne Shaphren, Randy Godwin, Ryan Hogan, Matthew Lieff, Perry McCarney, Nicholas Lovell, Mike Wever, Meg A. Write, Elsa Darcy, PIP Writer, Julie Hockenberry.