Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models

Anonymous ACL submission

Abstract

As Large Language Models (LLMs) have become more advanced, they have outpaced the ability to accurately evaluate them. Not only is finding data to adequately probe particular model properties difficult, but evaluating the correctness of a model's free-form generation alone is a challenge. To address this, many evaluations now rely on LLMs themselves to judge the quality of outputs from other LLMs, typically using a single large model like GPT-4. While this method has grown in popularity, it is costly, introduces intra-model bias, and in this work, we find that the largest models are often unnecessary. Instead, we propose evaluation with a Panel of LLm evaluators (PoLL) composed of a larger number of smaller models. Across three distinct judge settings and spanning six different datasets, we find that PoLL outperforms a single large judge, exhibits less intra-model bias due to its composition of disjoint model families, and does so while being over seven times less expensive.

1 Introduction

Evaluating generative language models is a challenging task; both in finding meaningful data to test the models, as well as evaluating the correctness of a generated response (Biderman et al., 2024). Multiple choice datasets like MMLU (Hendrycks et al., 2020) have become popular in part by side-stepping the difficulty of evaluating free-form generations. However, multiple-choice questions are in many ways probing a different property than that of a free-form generative task, which is typically closer to the desired use-case.

Many automatic metrics have been used across various tasks such as BLEU in machine translation (Papineni et al., 2002), ROUGE for summarization (Lin, 2004), and heuristic string match methods, such as normalized exact match (EM) and token level F1 for question answering (Rajpurkar et al., 2016). However, these simplistic methods commonly fail to analyze the intended property of interest. Question Answering (QA) metrics invariably lead to both false positive failures (e.g. superfluous token overlap) and more commonly, false negatives due to an incomplete set of gold reference answers (e.g. date format differences).

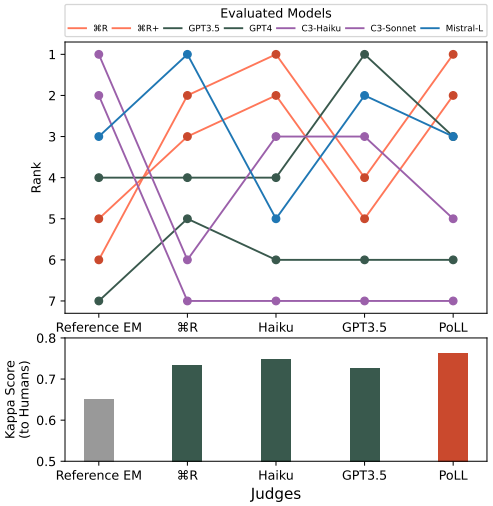

Figure 1: Top: Rankings of model performance change drastically depending on which LLM is used as the judge on KILT-NQ. Bottom: PoLLhas the highest Cohen's κ correlation with human judgements.

More recent methods attempt to address these issues by leveraging trained or prompted models as evaluators (Sellam et al., 2020; Zheng et al., 2024; Li et al., 2024c; Kocmi and Federmann, 2023a; Shen et al., 2023; Li et al., 2023). Prior work has shown that model-based scoring methods often correlate better with human judgements than heuristic metrics like EM (Bohnet et al., 2022; Zheng et al., 2024) and that strong evaluator models generalize well across different tasks (Huang et al., 2024).

Unfortunately, while the use of LLMs like GPT-4 as evaluators has become increasingly common, it has also been observed that evaluator models

tend to have their own biases; often recognizing and preferring their own outputs over those of other models (Panickssery et al., 2024). Additionally, it is most common to use the largest, most universally capable models as evaluators, which is both slow and costly, limiting applicability and access.

In this paper, we perform experiments across three settings (single-hop QA, multi-hop QA, and Chatbot Arena), spanning six datasets, and make the following contributions:

1. We propose to evaluate LLM generations using a Panel of LLm evaluators (PoLL) composed of different model families rather than a single large judge (Section 2).
2. We show that using an instantiation of PoLL correlates better with human judgements compared to a single large judge (GPT-4), and is over seven times cheaper (Sections 3.3).
3. In some scenarios, GPT-4 is a relatively weak judge, exhibiting high variance with minor changes to the prompt (Section 3.4).
4. Intra-model scoring bias is reduced by pooling judgements across a panel of heterogeneous evaluator models (Section 3.5).

2 Methods

2.1 Background: LLM as a Judge

A judge evaluator model J is used to score the output a from a test model A.

Single-point Scoring Here J rates the quality of a single model output independently of any point of comparison as score $= J(a)$ (e.g. Kocmi and Federmann (2023b)). J_{prompt} often includes natural language instructions on how to grade the generation (i.e. properties of good or bad outputs).

Reference-based Scoring The model is provided with a 'gold' reference r to compare with a and score $= J(a, r)$ (e.g. Zhu et al. (2023)). For example, in QA r would be a human-annotated 'correct' answer (See QA experiments in Section 3.1).

Pair-wise Scoring Lastly, the goal of pair-wise scoring is to choose which of two outputs (a and b) is better (e.g. Zheng et al. (2024)). Here score $= J(a, b)$[1] where the score's form can vary but is often a Likert scale (Likert, 1932) such as $a > b$, $a \approx b$, $a < b$ (See ChatBot Arena evaluations).

[1]To avoid ordering bias, it is common to test both $a - b$ and $b - a$ orderings as done in Section 3.1

2.2 Panel of LLM Evaluators

The above settings assume a single capable judge. However, as outlined earlier, a major issue with relying on a single model J, such as GPT-4, is the introduction of intra-model bias. To address this, we propose scoring answer correctness based not on a single judge, but instead on a panel composed of multiple evaluator models.

To calculate the PoLL score, each evaluator model independently scores an output a just as they would in the previous settings. Those individual scores are then pooled together through a voting function[2] such that the final score $= f(j \in P : j(a))$ where P is a panel composed of individual judges j and f is a voting function.

Pooling LLMs has also been explored previously (Jiang et al., 2023; Li et al., 2024a) and Li et al. (2023) proposed an evaluation method with multiple judges to reduce bias though only looked at large models on pair-wise evaluations. Similar pooling techniques reduce variance in human annotations by normalizing out both natural variation in human judgements caused by their own subjective biases as well as human error (Voorhees, 1998).

3 Experimental Results

3.1 Tasks and Datasets

Single-hop Question Answering Open-book QA tasks have models generate an answer conditioned on some retrieved evidence from a knowledge source (eg. internet or wikipedia) for the model to. Each question is accompanied by a human annotated reference answer. In our QA experiments, we include the commonly used exact match (EM) metric as a point of comparison. We experiment on KILT (Petroni et al., 2021) versions of Natural Questions (NQ), TriviaQA (TQA), and HotpotQA (HPQA) (Kwiatkowski et al., 2019; Joshi et al., 2017; Yang et al., 2018) (details in Section A.2.2).

Multi-hop Question Answering Multi-hop question answering is similar to the single-hop setting in Section 3.1, but the questions are designed such that models must perform multiple rounds of retrieval to answer sub-questions and collect sufficient evidence to ultimately answer the initial question. We perform experiments on two datasets: Bamboogle (Press et al., 2023) and HPQA. See Section A.2.3 for additional details.

[2]In this work we consider both max and average pool voting, though other functions could also work well.

Chatbot Arena Hard Chatbot Arena[3] is a popular benchmark for evaluating LLM head-to-head performance. This crowd-sourced effort has users prompt a pair of anonymized LLMs and rate which output is better (Zheng et al., 2024). Recently, Li et al. (2024b) mined a 'hard' subset of 500 prompts and had an LLM judge (GPT-4) score a model output against a baseline generation (a different version of GPT-4) in a head-to-head comparison.

At the time of writing, the repository only contained generated answers from GPT-3.5 and GPT-4. We therefore generated new outputs for all remaining models and subsequently scored generated answers with the various judges using the original codebase from Li et al. (2024b)[4]. The raw scores can be found in Table 7. Following the original work, we treat Chatbot Arena crowdsourced annotations as ground truth for calculating correlation between evaluator models and human judgements.

3.2 PoLL Composition and Voting

In our experiments, We construct a PoLL from three models drawn from three disparate model families (Command R, Haiku, and GPT-3.5). We consider two different voting functions for aggregating scores across the judges. For QA datasets, we use max voting, as all judgements are binary [correct, incorrect]. For Chatbot Arena we instead use average pooling because judgements are scores ranging from 1-5 and a three judge panel often does not produce a clear majority decision[5]. See Section A.1 for more details on the considered models.

Prompting Judges As alluded to in Section 2.1, judge models are prompted in different ways depending on the task. Our QA experiments use reference-based scoring and model prompts contain few-shot in-context examples of valid and invalid q, a, r triples. We also experiment with variations on these prompts in Section 3.4. Further details on judge prompts may be found in the appendix.

In our experiments on Chatbot Arena Hard, we follow the same pair-wise scoring setup of the original work. We do not modify the judge prompts in any way from the original implementation.

3.3 Correlation to Human Judgements

In Table 1 (left) we can see how the ratings from different evaluator judges, on different single-hop QA datasets from KILT, correlate with human judgements as measured by Cohen's κ. We see that overall, PoLL has the strongest correlation across various tasks, while GPT-4 is one of the weaker evaluators on this particular task setup (See 3.4 for further analysis and Section A.4 for annotation details). The right portion of the table shows correlation between judge model rankings and human judgement. Following Li et al. (2024b), we treat crowd-sourced ELO rankings from Chatbot Arena as ground truth. We calculate both Kendall Tau (Kendall, 1938) and Pearson Correlation (Pearson, 1895) of the ranked list produced by each of the judge methods with respect to this ground truth ranking. We find that PoLL is best correlated with the gold rankings, particularly at the top of the ranked list as shown in Figure 2.

Judge	NQ	TQA	HPQA	CBA-P	CBA-K
EM	0.651	0.827	0.662	-	-
GPT-4	0.627	0.841	0.830	0.817	0.667
CMD-R	0.734	0.902	0.815	0.817	0.676
Haiku	0.749	0.894	**0.873**	0.883	0.722
GPT-3.5	0.726	0.859	0.833	0.883	0.730
PoLL	**0.763**	**0.906**	0.867	**0.917**	**0.778**

Table 1: Judge-Human correlation where best results are indicated by the bold font and second best results are underlined. Left show single-hop QA datasets measured with Cohen's κ and right shows Pearson (CBA-P) and Kendall-Tau (CBA-K) correlations compared to the Chatbot Arena overall leaderboard.

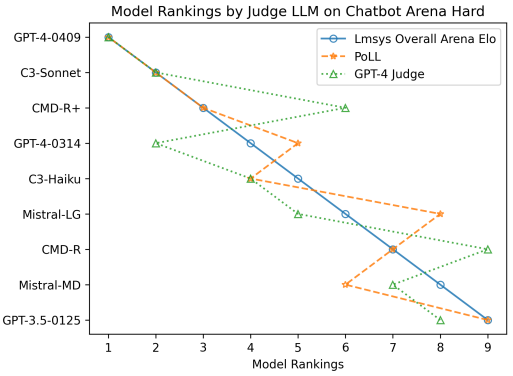

Figure 2: Rankings of model performance on Chatbot Arena Hard judged by GPT-4 or PoLL. Y-axis is model being judged, diagonal is original Chatbot Arena ranks.

3.4 Judgement Variance by Prompt Changes

Based on the observation that GPT-4 was the weakest judge model on our KILT evaluations, we investigated how the model reacts to modifications to its

[3] https://chat.lmsys.org/

[4] https://github.com/lm-sys/arena-hard

[5] Max voting with average voting fallback yielded similar scores and identical overall ranking.

Prompt Variant	Kappa
Zero-shot	0.518
Few-Shot Standard	0.627
+No Instruction Line	0.594
+Move Instruction Line	0.637
+Chat Formatted Shots	0.561
+'don't overthink'	0.725

Table 2: Kappa values on NQ for different prompt variants with GPT-4 as judge. For reference, GPT3.5 with the few-shot standard prompt achieves 0.726 (table 1).

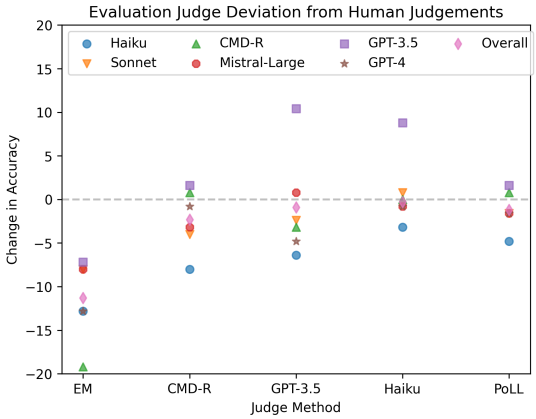

Figure 3: Accuracy changes of different evaluation judges compared to human judgements on Bamboogle.

prompt. GPT-4 is the most powerful judge model we tested, yet it performed worse than less capable models on what is essentially a fuzzy string matching exercise. We hypothesize that may be because GPT-4 is over-reasoning and injecting too much background knowledge into determining the correctness of an answer rather than simply aligning the gold reference with the generation. Different prompting strategies caused κ to vary between .518 and .725 with the most effective strategy is an explicit instruction to the model not to 'overthink' and not to concern itself with the wider factuality of the answers with respect to the outside world[6]. See Table 3 in Section A.2.1 for full results.

3.5 Judge Bias and Consistency

One of the main motivators for replacing a single large judge with a PoLL is to reduce bias in evaluation. To analyze this, we compared the delta in absolute scores for our individual judges and PoLL relative to human judgements on our multi-hop datasets. Figures 4 and 3 show judge scores for different models and how far those predictions deviate from human annotator decisions on HotPotQA and Bamboogle (the dotted line at 0). Overall, PoLL has the smallest spread in scores, with a standard deviation of 2.2, compared to EM and individual judges. GPT-3.5 has the highest spread, with a standard deviation of 6.1. We also see in Figure 3 that the highest positive delta for each individual model being scored occurs when it is judged by itself.

In Figure 2, we assessed model rankings produced on Chatbot Arena Hard by PoLL as compared to the GPT-4 judge from the original work. The 'gold' ranking appears on the diagonal, representing the rankings from the original Chatbot Arena ELO. We find that PoLL rankings correlate better with the ground truth, particularly at the top of the ranked list. We can clearly observe intra-model bias as the GPT-4 judge ranks another GPT-4 variant in position 2, higher than its actual position 4; in line with findings from prior works observing GPT-4's self-preference (Zheng et al., 2024; Panickssery et al., 2024)

3.6 Cost and Latency

At the time of writing, the cost of running our specific instance of PoLL is \$1.25/input[7] + \$4.25/output, whereas the cost of running GPT-4 Turbo is \$10/input + \$30/output. Depending on the ratio of input-to-output tokens in a given task, running the entire three model PoLL is seven to eight times less expensive than running a single GPT-4 judge.

We did not run explicit latency evaluations and many factors like model choice, serving platform, and more can impact speed. In general though, running a collection of smaller models in parallel (as in PoLL) is faster than a single big model.

4 Conclusions

In this paper, we showed how a Panel of LLM Evaluators composed of smaller models is an effective method for evaluating LLM performance, that shows high correlation with humans and reduces intra-model bias, latency, and cost. The benefits of PoLL are bolstered by the finding that there is not a single 'best' judge across all settings, while PoLL performs well consistently.

[6] the 'don't overthink' prompt had a negligible positive effect on Haiku, minor negative effect to Command R and a severe negative impact on GPT-3.5.

[7] per million tokens

5 Limitations

In this work, we investigated only three evaluator settings and a limited number of judges and panel compositions. While PoLL proved an effective alternative to a single large model in these settings, further work is needed to see how broadly applicable the method is. For example, math and reasoning settings where models often struggle (Zheng et al., 2024). We also leave the task of 'panel selection', or identifying the best models to include in PoLL in terms of quality and cost, to future work.

6 Ethical Impacts

Accurate evaluations are key to assessing the quality of models and steering them towards desired behaviors and outcomes. While PoLL attempts to reduce some form of bias in evaluations based on LLM judges, there will still be further biases that shift the models towards a particular generation. Additionally, access to accurate forms of evaluation is important for broader applicability. We hope that versions of PoLL composed entirely of fully open source models can be used by the community to perform accurate evaluations with lower costs.

References

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. 2024. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Hui Huang, Yingqi Qu, Jing Liu, Muyun Yang, and Tiejun Zhao. 2024. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers. *arXiv preprint arXiv:2403.02839*.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *ArXiv*, abs/2004.04906.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Tom Kocmi and Christian Federmann. 2023a. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024a. More agents is all you need. *arXiv preprint arXiv:2402.05120*.

Ruosen Li, Teerth Patel, and Xinya Du. 2023. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024b. From live data to high-quality benchmarks: The arena-hard pipeline.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024c. Leveraging large language models for nlg evaluation: A survey. *arXiv e-prints*, pages arXiv–2401.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *arXiv e-prints*, pages arXiv–2401.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Karl Pearson. 1895. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Ellen M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 315–323, New York, NY, USA. Association for Computing Machinery.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *ArXiv*, abs/2401.00368.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

A Appendix

A.1 Model Families

Command R Family: Command R (CMD-R, 35B) and Command R+ (CMD-R+, 103B) are open weight models created by Cohere[8]. We consider Command R as one of the models in the PoLL.
GPT Family: GPT-3.5 and GPT-4 (OpenAI et al., 2024) are two of the most widely used models. GPT-4 is typically the default choice for models as judges. We use GPT-3.5 as a member of our PoLL.
Claude-3 Family: Models built by Anthropic include (in increasing size) Haiku, Sonnet, and Opus (Anthropic, 2024). Our PoLLincludes Haiku.
Mistral Family: Mistral Large (Mistral-LG) and Mistral Medium (Mistral-MD) are proprietary models created by Mistral[9]. We did not conduct experiments with Mistral judges but evaluate their generations as a point of comparison to have a model 'unaffiliated' with any judges.

A.2 Additional Experimental Details

A.2.1 Judgement Variance by Prompt

In Table 3, we can see how the correlation between GPT-4 and human annotators varies as the prompt changes. In all cases, having in-context examples improves the performance over zero-shot and the most effective strategy is an explicit instruction to the model not to 'overthink' and not to concern itself with the wider factuality of the answers with respect to the outside world. These changes bring the agreement level for GPT-4 up to the level of GPT-3.5 when using our few-shot standard prompt, though still below Command-R, Haiku, and PoLL.

Prompt Variant	Kappa
Zero-shot	0.518
Few-Shot Standard	0.627
+No Instruction Line	0.594
+Move Instruction Line	0.637
+Chat Formatted Shots	0.561
+'don't overthink'	0.725

Table 3: Kappa values on NQ for different prompt variants with GPT-4 as judge. . For reference, GPT3.5 with the few-shot standard prompt achieves 0.726 (table 1)

[8]Release blog of Command R https://cohere.com/blog/command-r and Command R+ https://cohere.com/blog/command-r-plus-microsoft-azure

[9]Release blog of Mistral Large: https://mistral.ai/news/mistral-large/ and Mistral Medium: https://mistral.ai/news/la-plateforme/.

A.2.2 Single-hop QA

Metrics In addition to prompted judge scores, we also include the widely used exact match (EM) metric. We specifically use the 'containment' version of EM from prior work which is more amenable to LLM long-form generation (Liu et al., 2024) and checks if a reference answer string appears within the generated model response (after normalization).
Datasets KILT (Petroni et al., 2021) is a popular benchmark that measures a model's open domain question-answering capabilities with 1) a retrieval step to retrieve relevant documents from a knowledge source and 2) an answer generation step taking into account the relevant documents from (1). We select three popular test sets from the KILT benchmark, Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (TQA) (Joshi et al., 2017), and HotpotQA (HPQA) (Yang et al., 2018) for evaluation. The Natural Questions test set consists of real user questions sampled from a popular web search engine. TriviaQA is a test set with complex questions written by trivia enthusiasts. HotpotQA is a multi-hop question answering test set where questions require finding and reasoning over multiple supporting documents to answer. For all three test sets, the knowledge source is the corpus of Wikipedia documents form the KILT benchmark (Petroni et al., 2021).

KILT Experimental Details: For KILT, we use the DPR-formatted 100-word chunked KILT wikipedia dump (Karpukhin et al., 2020; Petroni et al., 2021).[10] Then, for all questions, we retrieve and cache the top-10 snippets from the dump, using the GTE-large dense retrieval model (Wang et al., 2023), ready to be used to generate answers from different models.

Generating answers for judgement: To generate answers from all considered models, we ask the question as an unmodified chat message and pass in the snippets using the dedicated `documents` parameter of the model family's API where available (Command-R Family)[11]. For model families that do no have a specific `documents` api parameter (GPT, we instead adopt the question answering

[10]The pre-chunked dump can be downloaded here: https://github.com/facebookresearch/DPR/issues/186#issuecomment-923268677

[11]The prompt template that gets used when using this parameter is described at https://huggingface.co/CohereForAI/c4ai-command-r-v01 with additional details at https://docs.cohere.com/docs/prompting-command-r

prompt template used in Liu et al. (2024), which is shown in table 9.

Judge Prompts We use the same prompt-template structure for KILT judgements as for Multihop judgements. We initially observed relatively-lower levels of human-agreement when using the multihop prompt verbatim applied to KILT results experiments. We hypothesise this was due to mismatch in model-generated answer styles in the few-shot examples. We therefore create new judgement prompts the following procedure for each of the KILT datasets, randomly sampling a set of labelled fewshot examples from our pool human-labelled examples according to the following constraints:

- at least 1 annotated positive example from each model family's generations
- at least 1 annotated negative example from each model family's generations
- at least 1 example where the human disagrees with keyword-based exact match score.

We sweep over `n_shots` in $\{8, 10, 16\}$, picking the prompt that has strongest agreement on average across all judges for a small held-out portion of human labels. The resuling judgement prompts used for KILT in our experiments in the main paper are listed in tables 12, 11 and 13.

GPT-4 judge prompt ablation Here we include some additional details on prompt ablation for KILT NaturalQuestions mentioned in section 3.4:

- **Zero-shot**: the natural language instruction is supplied as system call. We then directly ask the model to judge with no labelled examples
- **Few-shot Standard**: This is the prompt in table 11 which is used for all other model judges.
- **No instruction Line**: Here we remove the natural language instruction from the Few-shot standard prompt, on the hypothesis the instruction is confusing the model. This hypothesis turns out to be false as agreement actually drops. (-0.03 $\Delta\kappa$)
- **Move instruction Line**: Here we modify the standard fewshot prompt by moving the instruction line into a separate system call. This results in a small improvement (+0.01 $\Delta\kappa$)
- **Chat-formatted shots**: Here we modify the standard fewshot prompt by formatting each fewshot example as a conversational turn between the user and the assistant. This reduces performance (-0.07 $\Delta\kappa$)
- **'don't overthink'**: Here, we replace the instruction line from the standard prompt's `You will be given a Question and a Provided Answer. Judge whether the Provided Answer is correct by comparing it to the Reference Answer. Differently formatted dates, people with missing middle names, and alternative spellings should all be considered the same. If the Provided Answer is correct say exactly "True", otherwise say "False".` to a wording which is intended to encourage the model to perform a simpler function and not incorporate external knowledge: `"You are judging whether a model has generated a correct answer to a question. Study the examples the user gives you as they will be very informative for how to do the task. The Reference Answers you get will be short. An model's answer will be longer, and can be considered correct if it contains the semantic content of short reference answer somewhere within it. Don't worry about factuality with respect to the real world, just judge the example based on what you see. No need to overthink this task, it really comes down to just soft matching.` This improves results for GPT-4 by +0.07 $\Delta\kappa$. Additional small surface level changes and moving the instruction to a system call lead to an additional +0.03 $\Delta\kappa$. The final optimized prompt for GPT-4 can be found in table 14.

The agreements for different models using the prompt optimized for GPT-4 can be found in table 4.

A.2.3 Multihop

Bamboogle (Press et al., 2023) is 125 questions adversarially constructed to require multiple internet searches to successfully answer. **HotpotQA** asks questions meant to be answered with an accompanying wikipedia dump (we consider the subset used in (Shinn et al., 2024)).

	Standard	'don't overthink'
EM	0.651	0.651
GPT4	0.627	0.725
CMD-R	0.734	0.687
Haiku	0.749	0.757
GPT-3.5	0.729	0.509
PoLL	**0.763**	0.699

Table 4: Kappa values for KILT NQ for the standard prompt and the prompt optimized for GPT-4 (note that the EM baseline is not LLM-dependent and therefore not dependent on prompt)

Generating answers for judgement – React Agents: All models follow the basic REACT setup for tool-use agents which incorporate a Thought-Action-Observation loop to iteratively call tools and reason over their outputs to find a final answer (Yao et al., 2022). We use LangChain tool-use implementations for each model when available. On Bamboogle, agents have access to an internet search tool and on HotPotQA, agents have access to a search tool over wikipedia that does dense embedding search using Cohere EmbedV3 + Rerank 3. All search tool calls return the top two most relevant documents to the model.

Judge Prompt The prompt used for judging multihop answers can be found in table 10

A.3 Additional Results

A.3.1 Multi-hop QA

In Tables 5 and 6 we can see κ correlations on Bamboogle and HotPotQA multi-hop question answering.

	EM	GPT-3.5	Cmd-R	Haiku	PoLL
Haiku	0.746	0.772	0.741	0.802	**0.803**
Sonnet	0.767	0.879	0.880	0.877	**0.896**
CMD-R	0.599	0.881	0.896	0.897	**0.938**
Mistral-LG	0.841	0.760	**0.936**	**0.936**	0.872
GPT-3.5	0.849	0.886	**0.935**	**0.935**	**0.935**
GPT-4	0.677	0.862	0.877	0.877	**0.903**
Total	0.762	0.846	0.885	**0.898**	0.896

Table 5: Kappa values for various chatbot models on Bamboogle

A.3.2 Arena Hard Scores

Raw scores for Arena-hard coming from PoLL are shown in Table 7.

	EM	GPT-3.5	Cmd-R	Haiku	PoLL
Haiku	0.703	**0.910**	0.890	0.864	0.889
Sonnet	0.715	0.796	**0.847**	0.844	0.820
CMD-R	0.780	0.815	**0.909**	0.906	0.859
Mistral-LG	0.788	0.754	**0.895**	0.894	0.853
GPT-3.5	0.799	0.757	**0.879**	**0.879**	0.858
GPT-4	0.768	**0.890**	0.835	0.835	**0.890**
Total	0.773	0.827	**0.886**	0.849	0.871

Table 6: Kappa values for various chatbot models on multihop HotPotQA

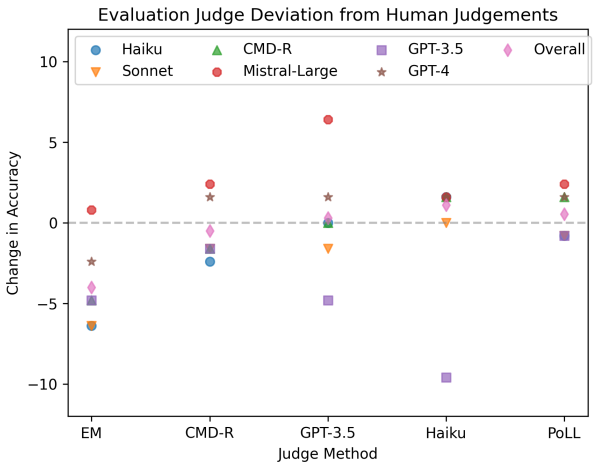

Figure 4: Accuracy changes of different evaluation judges as compared to human judgements on HotPotQA (multi-hop).

A.4 Human Annotations

To gather human reference judgements, we utilized a professional highly-qualified annotation workforce. Annotators were shown a single anonymized model generated answer at a time along with the original question and reference answer. Annotators were asked to judge whether the reference answer is semantically contained inside the generated answer[12].

Annotations were performed by profession annotators with diverse demographic and professional backgrounds including novelists, copywriters, copy editors, and journalists. Single-hop TriviaQA, Bamboogle, and multi-hop HotPotQA were singly annotated while Natural Questions and single-hop HotPotQA are triple annotated. Experiments on triple annotated datasets were based on a pooled super-annotator decision based on majority voting. The total number of annotations is shown in Table 8.

[12]By asking for answer correctness alone, we hope to minimize any potential annotator bias or preference towards a particular model's style.

Model	Score	95% CI
GPT-4-turbo	68.7	(-2.1, 2.2)
Sonnet	57.6	(-3.2, 2.2)
CMD-R+	57.1	(-2.9, 2.5)
Haiku	55.9	(-2.1, 2.2)
GPT-4-0314	50.0	(0.0, 0.0)
Mistral-MD	49.0	(-2.3, 2.7)
CMD-R	48.0	(-1.9, 2.7)
Mistral-LG	43.7	(-1.9, 2.2)
GPT-3.5	41.2	(-2.1, 2.5)

Table 7: Scores from Arena Hard as scored by the PoLL using average pooling.

Table 8: Total Number of Human Annotations by Dataset

Dataset	Annotated Examples	Annotations per Example
Natural Questions	688	Triple
TriviaQA	316	Single
HotPotQA (single)	784	Triple
HotPotQA (multi)	595	Single
Bamboogle	750	Single

```
Write a high-quality answer for the given question using only the provided search results (some of
which might be irrelevant).

Document [1](Title: Lake Eyre basin) keeping pace with evaporation. In contrast, the flow of the
Mississippi could fill Lake Eyre in 22 days, that of the Amazon in just 3 days. Other lakes in the
basin include Lake Frome, Lake Yamma Yamma and Lake Hart. Geography Rivers. The Cooper Creek, Finke
River, Georgina River and Diamantina River are the four main rivers of the basin. Other desert rivers
include the Hale River, Plenty River and Todd River that flow from the south east of the Northern
Territory, south. In the
Document [2](Title: Lake Eyre basin) make it as far south as Lake Eyre, although the story is told
that this happened once early in the 20th century. In extreme events, water from the Finke River flows
into the Macumba River, which empties into Lake Eyre, a total distance from headwater streams of
around . Major tributaries include Ellery Creek, and the Palmer and Hugh Rivers. The Georgina River
system originates on the Barkly Tableland, near the Northern Territory-Queensland border, north-west
of Mount Isa and not far south of the Gulf of

...

Document [10](Title: Lake Eyre) from the north-east part of the Lake Eyre Basin—in outback
(south-west and central) Queensland—flow towards the lake through the Channel Country. The amount of
water from the monsoon determines whether water will reach the lake and, if it does, how deep the lake
will get. The average rainfall in the area of the lake is per year. The altitude usually attributed to
Kati Thanda–Lake Eyre refers to the deepest parts of the lake floor, in Belt Bay and the Madigan Gulf

Question: where does the water come from to fill lake eyre
Answer:
```

Table 9: Example of a single-hop question answering prompt from Liu et al. (2024) used for KILT answer generations where LLM API does not have a documents parameter.

```
You will be given a Question and a Provided Answer. Judge whether the Provided Answer is correct by comparing it to the Reference Answer. Differently formatted dates, people with missing middle names, and alternative spellings should all be considered the same. If the Provided Answer is correct say exactly "True", otherwise say "False".

Question 1: "When did the president who set the precedent of a two term limit leave office?"
Provided Answer: "George Washington set the precedent of a two-term limit when he decided not to seek a third term in 1796. He left office in 4 March, 1797."
Reference Answer: "March 4, 1797"
Correct: True


Question 2: "Where does Śivarāma Swami conduct courses on Vaishnava Theology?"
Provided Answer: "Śivarāma Swami conducts courses on Vaishnava Theology at Bhaktivedanta Manor."
Reference Answer: "Where does Śivarāma Swami conduct courses on Vaishnava Theology?"
Correct: False


Question 3: "The most populous city in Punjab is how large (area wise)?"
Provided Answer: "Ludhiana, the most populous city in Punjab, covers an area of 3,767 square kilometres according to the website of Punjab's Department of Agriculture & Farmer Welfare. Another source states it is 310 square kilometres, making it the most densely populated urban centre in the state."
Reference Answer: "310 square kilometers"
Correct: True


Question 4: "Who was mayor of New York City when Fiorello H. La Guardia was born?" Provided Answer: "Frank Edson was mayor of New York City when Fiorello H. La Guardia was born on December 11, 1882. Edson served as mayor from 1881 to 1882, and La Guardia himself later became the 99th Mayor of New York City, serving from 1934 to 1946 or 1945."
Reference Answer: "William Russell Grace"
Correct: False


Question 5: "What is the third letter of the top level domain of the military?"
Provided Answer: "The third letter of the top-level domain for the military, '.mil', is 'l'.
The domain name 'mil' is derived from the word 'military'. It is the only top-level domain for a country's military, a legacy of the US military's role in the creation of the internet."
Reference Answer: "l"
Correct: True


Question 6: "Ed Oxenbould plays Tyler in a film directed by which famous director?"
Provided Answer: "Ed Oxenbould played Tyler in *The Visit*"
Reference Answer: "M. Night Shyamalan"
Correct: False


Question 7: "Alice David is the voice of Lara Croft in a video game developed by which company ?"
Provided Answer: "Alice David is the voice of Lara Croft in the French dub of the video game *Tomb Raider*."
Reference Answer: "Crystal Dynamics"
Correct: False


Question 8: "{QUESTION}"
Provided Answer: "{GEN_ANSWER}"
Reference Answer: "{GOLD_ANSWER}"
Correct:
```

Table 10: Multihop Judge Prompt

```
You will be given a Question and a Provided Answer. Judge whether the Provided Answer is correct by comparing it to the Reference Answer. Differently formatted dates, people with missing middle
names, and alternative spellings should all be considered the same. If the Provided Answer is correct say exactly "True", otherwise say "False".

Question 1: "when did the nfl start playing in london"
Provided Answer: "According to the provided search results, the NFL started playing regular season games in London as part of the NFL International Series in 2007. Specifically:

Document [5] states: "The NFL International Series was inaugurated in 2007 to host NFL regular season games outside the United States. Played at the new Wembley Stadium in London (rebuilt and
reopened in 2007), the series increased from one to two games for the 2013 season, to three games for the 2014 season, and then to four games from the 2017 season."

Document [9] also mentions: "Since 2007, the league has held multiple regular season games in London each season as part of NFL London Games, allowing the league to test solutions to some of the
challenges facing a hypothetical London franchise."

So 2007 marked the start of the NFL playing regular season games at the newly rebuilt Wembley Stadium in London as part of the International Series. Prior to that, there were some preseason games
played in London, but 2007 was the first year for regular season NFL games in the city."
Reference Answer: "since 2007"
Correct: True


Question 2: "where is wind power used most in the united states"
Provided Answer: "Iowa is the state that generates the highest percentage of energy from wind power, with more than 35% of its electric power coming from wind. Alternatively, North Dakota has the
highest per capita wind generation.

The state with the most installed wind power capacity is Texas, followed by Iowa. Texas's capacity is around 20,000 MW, almost double that of Iowa's 3,670 MW. The Roscoe Wind Farm in Texas is also
the world's largest wind farm at 780 MW.

California has doubled its wind power capacity since 2002 and now gets about 5% of its energy from wind power. This is still lower than the twelve U.S. states that get more than 20% of their energy
from wind."
Reference Answer: "Texas"
Correct: True


Question 3: "who is the owner of phoenix mall in chennai"
Provided Answer: "The owner of Phoenix Marketcity in Chennai is Phoenix Mills Limited, as mentioned in the document titled "Phoenix Marketcity (Chennai)." This mall was jointly developed by Phoenix
Mills Limited and Crest Ventures Ltd."
Reference Answer: "Phoenix Mills Limited"
Correct: True


Question 4: "number of degree of freedom for plane mechanism"
Provided Answer: "In mechanics, the degree of freedom (DOF) of a mechanical system is the number of independent parameters that define its configuration. The number of degrees of freedom for a
mechanism on a two-dimensional plane is three, which consists of two translations and one rotation.

For a system of multiple bodies, the combined DOF is the sum of the DOFs of the individual bodies, minus the internal constraints on their relative motion. Therefore, a mechanism formed of multiple
rigid bodies may have more than three DOFs."
Reference Answer: "two"
Correct: False


Question 5: "who had the best strike rate among batsmen who scored over 200 runs in the world cup"
Provided Answer: "The provided documents do not specify the strike rate of all batsmen who scored over 200 runs in the World Cup. However, among the batsmen mentioned who have scored over 200 runs in
the World Cup, Jehan Mubarak has the highest ever strike rate in a T20 World Cup match for Sri Lanka, with a strike rate of 353.84 (Document [6]). But it's important to note that this strike rate is
not necessarily reflective of his overall performance in the World Cup, as it could have been achieved in a single match.
For a more accurate comparison of strike rates among batsmen who scored over 200 runs in the World Cup, a comprehensive list of all such batsmen and their respective strike rates would be required."
Reference Answer: "Rohit Sharma"
Correct: False


Question 6: "who played ben stone son on law and order"
Provided Answer: "According to the provided search results, Ben Stone's son Peter Stone was played by Philip Winchester. Some key details:
- Peter Stone is the son of Benjamin "Ben" Stone, the former Executive Assistant District Attorney portrayed by Michael Moriarty on Law & Order.
- Peter first appeared as a character on the Chicago P.D. episode "Justice", which served as a backdoor pilot for the short-lived spin-off Chicago Justice, where he was the lead character played by
Philip Winchester.
- After Chicago Justice was canceled, Philip Winchester reprised his role as Peter Stone in Season 19 of Law & Order: Special Victims Unit, where he became the Assistant District Attorney for the Sex
Crimes Bureau.
- In several SVU episodes, Peter describes having a strained relationship with his workaholic father Ben Stone, who neglected his family in favor of his prosecutorial duties.
So in summary, Philip Winchester played the role of Peter Stone, the son of the iconic Law & Order character Ben Stone, across the shows Chicago Justice and Law & Order: SVU."
Reference Answer: "Michael Moriarty"
Correct: False


Question 7: "who won so you think you can dance season 13"
Provided Answer: "I'm sorry, I couldn't find any information about the winner of season 13 of So You Think You Can Dance." Reference Answer: "Leon " Kida " Burns"
Correct: False


Question 8: "when did sweet caroline start at red sox games"
Provided Answer: "The song "Sweet Caroline" started being played at Fenway Park, home of the Boston Red Sox, since at least 1997. However, it has been played consistently in the middle of the eighth
inning at every game since 2002."
Reference Answer: "at least 1997"
Correct: True


Question 9: "when did end of the road come out"
Provided Answer: "I found multiple results for "End of the Road":
## End of the Road (1944 film)
End of the Road is a 1944 American crime film directed by George Blair and written by Denison Clift, Gertrude Walker and Albert Beich. The film stars Edward Norris, John Abbott, June Storey, Jonathan
Hale, Pierre Watkin and Ted Hecht.
## End of the Road (1970 film)
End of the Road is a 1970 American comedy-drama film directed, co-written, and edited by Aram Avakian and adapted from a 1958 novel by John Barth. The film stars Stacy Keach, James Earl Jones and
Harris Yulin.
## End of the Road (Boyz II Men song)
End of the Road is a single recorded by American R&B group Boyz II Men for the "Boomerang" soundtrack. It was released on June 30, 1992.

## End of the Road (Jerry Lee Lewis song)
End of the Road is a song by Jerry Lee Lewis. It appeared on the 1976 Sun Records compilation album "The Original Jerry Lee Lewis" which was released by the Sun Record Company.
Which of these were you referring to?"
Reference Answer: "June 30 , 1992"
Correct: True


Question 10: "who played john connor in the original terminator"
Provided Answer: "John Connor is not physically portrayed by any actor in the original "Terminator" film released in 1984, as he is only mentioned and not seen. The character of John Connor is first
referred to in this film as the future leader of the human resistance against the machines, but he does not make an appearance."
Reference Answer: "Edward Furlong throughout the remainder of the film"
Correct: False


Question 11: "{QUESTION}"
Provided Answer: "{GEN_ANSWER}"
Reference Answer: "{GOLD_ANSWER}"
Correct:
```

Table 11: KILT NaturalQuestions judge prompt.

You will be given a Question and a Provided Answer. Judge whether the Provided Answer is correct by comparing it to the Reference Answer. Differently formatted dates, people with missing middle names, and alternative spellings should all be considered the same. If the Provided Answer is correct say exactly "True", otherwise say "False".

Question 1: "Who was also an actor, Serri or John Fogerty?"
Provided Answer: "Based on the provided search results, there is no information indicating that John Fogerty has pursued acting as a profession or has been involved in acting projects. The search results focus on his music career, including his time with Creedence Clearwater Revival, his solo career, and various performances and recognitions he has received over the years. Therefore, without information on Serri in the provided documents, it cannot be determined if Serri was an actor. However, it is clear from the search results that John Fogerty is primarily known as a musician, singer, and songwriter, not as an actor."
Reference Answer: "Serri"
Correct: False

Question 2: "The 37th Disney animated feature film Tarzan features You'll be in my heart by which artist?" Provided Answer: "The 37th Disney animated feature film Tarzan's soundtrack, Tarzan: An Original Walt Disney Records, includes the song You'll Be in My Heart written and performed by Phil Collins. The song won the Academy Award for Best Original Song and the Golden Globe Award for Best Original Song. Collins was originally hired as a songwriter but ended up performing the song as well."
Reference Answer: "Phil Collins"
Correct: True

Question 3: "Who is the current drummer of the band who did the song "What Lovers Do"?"
Provided Answer: "Unfortunately, none of the provided search results directly mention the current drummer for the band Maroon 5 who performed the song "What Lovers Do". The results provide background information on the song itself, its music video, chart performance, and other details, but do not specify the band members or drummer.

Maroon 5 is a well-known pop rock band fronted by Adam Levine, but their full current lineup and specific drummer for the "What Lovers Do" era is not stated in these search results. To definitively answer who the drummer was, additional information directly about Maroon 5ś band members around 2017 when "What Lovers Do" was released would be needed." Reference Answer: "Matt Flynn"
Correct: False

Question 4: "Who is older, Ferdi Taygan or Mahesh Bhupathi?"
Provided Answer: "Unfortunately, none of the provided search results contain any information about Ferdi Taygan or their age to compare with Mahesh Bhupathi's age. The search results are all focused on providing biographical details about Mahesh Bhupathi, an Indian professional tennis player, and some other individuals named Mahesh Bhatt who appear to be unrelated to the question. Without any information about Ferdi Taygan, it is impossible to determine who is older between them and Mahesh Bhupathi based solely on these search results."
Reference Answer: "Ferdi Taygan"
Correct: False

Question 5: "What is the birthday of the actress who was the Duchess in "The Revengers Tragedy"?"
Provided Answer: "Diana Quick, who played the Duchess in "The Revengers Tragedy", was born on 23rd September 1934."
Reference Answer: "23 November 1946"
Correct: False

Question 6: "Rejuvelac is kind of grain water invented and promoted by a "holistic health" practitioner born in which year ?"
Provided Answer: "Rejuvelac is a kind of grain water invented and promoted by Ann Wigmore, who was born in 1909."
Reference Answer: "1909"
Correct: True

Question 7: "Who is the current drummer of the band who did the song "What Lovers Do"?"
Provided Answer: "The song "What Lovers Do" is by the American pop rock band Maroon 5. The current drummer of Maroon 5 is Matt Flynn."
Reference Answer: "Matt Flynn"
Correct: True

Question 8: "The 37th Disney animated feature film Tarzan features You'll be in my heart by which artist?"
Provided Answer: "The song "Youll Be in My Heart" featured in the 1999 Disney animated film Tarzan was written and performed by Phil Collins.
The key details that support this answer from the provided search results:
1. Document [2] states: "Youll Be in My Heart" is a song by Phil Collins, from the 1999 Disney animated feature "Tarzan"."

2. Document [3] mentions: "For his contribution to the soundtrack, Collins received an American Music Award for Favorite Adult Contemporary Artist."

3. Document [5] says: "Collins was originally hired by Disney as a songwriter, but was chosen because he was a drummer at first and given his history as a drummer for the rock band Genesis, he appealed to the Disney production crew, who "wanted a strong jungle beat to accompany Tarzanś adventures". This "ballad" is one of five original songs he wrote for the film, and he ended up also performing this song as well."

4. Document [7] states: "The song is about "how love is a bond that cannot be broken"."

So in summary, the artist who wrote and performed the song "Youll Be in My Heart" featured in the Disney animated film Tarzan is Phil Collins."
Reference Answer: "Phil Collins"
Correct: True

Question 9: "How long is the river for which Frenchmans Creek is a tributary?"
Provided Answer: "The length of the river for which Frenchmans Creek is a tributary cannot be directly determined from the provided documents as they only provide information about the Frenchmans Creek itself and not the river it flows into. However, we can see that Frenchmans Creek is a tributary of the Darling River in New South Wales, Australia (Document 2), the Mississippi River in Missouri, USA (Document 3), and the Milk River in Saskatchewan, Canada, and Montana, USA (Document 5). The length of these rivers is not provided in the documents."
Reference Answer: "2844 km"
Correct: False

Question 10: "Conrad Anker located the body of a mountaineer who took part it the first three British expeditions to which mountain?"
Provided Answer: "Conrad Anker located the body of George Mallory on Mount Everest. Mallory was a British climber who took part in the first three British expeditions to Mount Everest, including the 1924 expedition during which he and his climbing partner Andrew Irvine disappeared while attempting to reach the summit. Anker found Mallory's body in 1999 as a member of a search team looking for the remains of the British climber.
References: [1] Conrad Anker [2] 2007 Altitude Everest expedition [9] Timeline of Mount Everest expeditions"
Reference Answer: "Mount Everest"
Correct: True

Question 11: "The city that contains the Yunnan Provincial Museum is also known by what nickname?"
Provided Answer: "The city that contains the Yunnan Provincial Museum, Kunming, is also known by the nickname "Spring City.""
Reference Answer: "Spring city"
Correct: True

Question 12: "The 1st Virginia Regiment saw action at a battle that led to the departure of what Royal Governor?" Provided Answer: "Based on the provided search results, the 1st Virginia Regiment saw action at the Battle of Great Bridge in December 1775, which led to the departure of Lord Dunmore, the Royal Governor of Virginia at the time.

Document [1] mentions that the 1st Virginia Regiment participated in the Battle of Great Bridge. Document [4] states that after the Battle of Great Bridge, "Dunmore responded by bombarding Norfolk with his ships on January 1, 1776." This implies that the Battle of Great Bridge, in which the 1st Virginia Regiment fought, led to Lord Dunmoreś departure from Virginia as the Royal Governor.

So the 1st Virginia Regimentś involvement in the Battle of Great Bridge resulted in the departure of Lord Dunmore, the Royal Governor of the Colony of Virginia during the American Revolutionary War."
Reference Answer: "Royal Governor Lord Dunmore"
Correct: True
Question 13: "The voice of Homer Simpson also voices what alien character in the television series "The Simpsons"?"
Provided Answer: "Dan Castellaneta voices Homer Simpson, Grampa Simpson, Krusty the Clown, Groundskeeper Willie, Mayor Quimby, Barney Gumble, and other adult characters."
Reference Answer: "Kodos"
Correct: False

Question 14: "Who were the producers of the show for which Chris McKay made his feature film directorial debut?" Provided Answer: "Chris McKay made his feature film directorial debut with "The Lego Batman Movie" in 2017. The producers of "The Lego Batman Movie" were not explicitly mentioned in the provided documents. However, it is noted that Seth Grahame-Smith co-wrote the screenplay for the movie, and Warner Bros. was involved in its development, as mentioned in Document 8. Additionally, Document 9 mentions that McKay had signed a first look deal for being a producer with Warner Bros. around the time of the movieś development. While this information does not directly answer the question regarding the specific producers of "The Lego Batman Movie," it does highlight the key individuals and entities involved in its creation and development." Reference Answer: "Dan Lin, Roy Lee, Phil Lord and Christopher Miller" Correct: False

Question 15: "Marion Greene was a health policy analyst for St. Judt Medical company, which had how many principal operations worldwide?"
Provided Answer: "The provided documents do not specify the number of principal operations St. Jude Medical company had worldwide. Document 1 mentions that Marion Greene was a health policy analyst for St. Jude Medical, but it does not provide details about the company's operations. Document 2 also mentions Greene's work for St. Jude Medical, but again, no information is given about the company's operations. The other documents do not pertain to St. Jude Medical. Therefore, I cannot provide an answer to this question based on the given documents."
Reference Answer: "20"
Correct: False
Question 16: "What is the name of the actress who has received a Drama Desk award and was part of the cast album for the 2007 Broadway production of the hit musical, "Grease"?"
Provided Answer: "Laura Osnes, who played the role of Sandy Dumbrowski in the 2007 Broadway production of "Grease", is a Drama Desk award winner and is featured on the cast album for the production."
Reference Answer: "Laura Ann Osnes"
Correct: True

Question 17: "{QUESTION}"
Provided Answer: "{GEN_ANSWER}"
Reference Answer: "{GOLD_ANSWER}"
Correct:

Table 12: KILT HotpotQA Judge prompt

```
You will be given a Question and a Provided Answer. Judge whether the Provided Answer is correct by comparing it to the Reference Answer. Differently formatted dates, people with missing middle names, and alternative spellings should all be considered the same. If the Provided Answer is correct say exactly "True", otherwise say "False".

Question 1: "If you landed at 'Santa Catarina Airport on which Island would you be in?"
Provided Answer: "If you landed at Santa Catarina Airport, you would be on the island of Santa Catarina, in the Brazilian state of the same name. The island is approximately 54km long and 18km wide and is home to the state capital, Florianópolis. It is the largest island in an archipelago of over 30 islands."
Reference Answer: "Geography of the Madeira Islands"
Correct: False


Question 2: "When it was a first performed at the Abbey Theatre in Dublin in 1907, which play caused rioting in the streets because of objections to the portrayal of rural Irish Catholic immorality?"
Provided Answer: "Based on the provided search results, the play that caused rioting in the streets when it was first performed at the Abbey Theatre in Dublin in 1907 due to objections over the portrayal of rural Irish Catholic immorality was "The Playboy of the Western World" by J.M. Synge.

Several documents mention the riots that erupted during and after the opening performance of Syngeś "The Playboy of the Western World" at the Abbey Theatre in January 1907. The riots were instigated by Irish nationalists and republicans who viewed the playś content as offensive to public morals and an insult to Ireland (Documents 1, 3, 7).

Document 5 specifically states that "the first performance of J.M. Syngeś play Ťhe Playboy of the Western Worldát the Abbey Theatre in Dublin triggers a week of rioting" in 1907.
Document 9 also notes that a performance of the play on January 28, 1907 "was interrupted by the audience who continue to boo, hiss and shout."

The objections seem to have stemmed from the playś depiction of patricide and perceived immorality in rural Irish Catholic society, which nationalists found offensive (Document 7). So "The Playboy of the Western World" by J.M. Synge was the controversial play that provoked riots when first staged at the Abbey Theatre in 1907 due to its portrayal of the Irish countryside."
Reference Answer: "'THE PLAYBOY OF THE WESTERN WORLD'"
Correct: True


Question 3: "The religious order of Poor Ladies of San Damiano has what more common name?" Provided Answer: "The religious order of Poor Ladies of San Damiano is more commonly known as the Poor Clares. This order was founded by Saints Clare of Assisi and Francis of Assisi in 1212, and it was the second Franciscan Order to be established. The Poor Clares were initially referred to as the Order of Poor Ladies, and later the Clarisses, the Minoresses, the Franciscan Clarist Order, and the Second Order of Saint Francis. However, the name Poor Clares is the most recognized and widely used today."
Reference Answer: "Franciscan Clarist Order"
Correct: False


Question 4: "Which country defeated the Ivory Coast to win football's Africa Cup of Nations?"
Provided Answer: "Based on the provided search results, Ghana defeated the Ivory Coast in the final of the 1992 Africa Cup of Nations. The key evidence is from Document 1 and Document 10:

Document 1 states: "Ivory Coast set a record for the competition by holding each of their opponents scoreless in the six matches of the final tournament after his contributions helped Ghana reach the final; he was, however, suspended for that match and Ghana lost to Ivory Coast in a penalty shootout that saw each side make 11 attempts to determine the winner."

Document 10 directly mentions: "The 1992 Africa Cup of Nations Final was a football match that took place on January 26, 1992, at the Stade de lÂmitié in Dakar, Senegal, to determine the winner of the 1992 African Cup of Nations. Ivory Coast defeated Ghana 11–10 on penalties after a goalless draw to win their first African Cup."

So the Ivory Coast defeated Ghana in the 1992 Africa Cup of Nations final after a penalty shootout. None of the other provided documents contradict this or indicate another country defeated the Ivory Coast to win that specific tournament."
Reference Answer: "health care in zambia"
Correct: False


Question 5: "Fred Scuttle was one of the best-known characters of which British comedian?"
Provided Answer: "The provided documents do not directly mention Fred Scuttle or the British comedian associated with this character. Therefore, based on the information available in the documents, it is not possible to accurately answer the question about which British comedian is best known for the character Fred Scuttle."
Reference Answer: "Alfred Hawthorne Hill"
Correct: False


Question 6: "What element is the main constituent of anthracite?"
Provided Answer: "The main constituent of anthracite is carbon. High-grade anthracite, for example, has a chemical formula represented by CHONS, indicating it contains carbon, hydrogen, oxygen, nitrogen, and sulfur, with carbon making up 94% of its composition. Ultra high-grade (UHG) anthracite typically has a minimum carbon content of 95%, further emphasizing carbon as the primary element in anthracite."
Reference Answer: "Carbon"
Correct: True


Question 7: "Which American played in his 10th Ryder Cup in 2014?"
Provided Answer: "Mickelson extended his own U.S. team record with his 10th appearance at the 2014 Ryder Cup."
Reference Answer: "philip alfred mickelson"
Correct: True


Question 8: "Who is the lead singer of the Sheffield band the Arctic Monkeys?"
Provided Answer: "The lead singer of the Sheffield band Arctic Monkeys is Alex Turner. Turner also plays guitar and is the frontman of the four-piece indie rock band.
Arctic Monkeys have released six studio albums, including a debut in 2002 titled "Whatever People Say I Am, That's What I'm Not"."
Reference Answer: "Alex Turner"
Correct: True


Question 9: "{QUESTION}"
Provided Answer: "{GEN_ANSWER}"
Reference Answer: "{GOLD_ANSWER}"
Correct:
```

Table 13: KILT TriviaQA Judge Prompt

```
SYSTEM: You are judging whether a model has generated a correct answer to a question. Study the examples the user gives you as they will be very informative for how to do the task. The Reference Answers you get will be short. An model's answer will be longer, and can be considered correct if it contains the semantic content of short reference answer somewhere within it. Don't worry about factuality with respect to the real world, just judge the example based on what you see. No need to overthink this task, it really comes down to just soft matching. Answer with only the word 'Yes' or 'No'

USER: Question: when did the nfl start playing in london Model Answer: According to the provided search results, the NFL started playing regular season games in London as part of the NFL International Series in 2007. Specifically:

Document [5] states: "The NFL International Series was inaugurated in 2007 to host NFL regular season games outside the United States. Played at the new Wembley Stadium in London (rebuilt and reopened in 2007), the series increased from one to two games for the 2013 season, to three games for the 2014 season, and then to four games from the 2017 season."

Document [9] also mentions: "Since 2007, the league has held multiple regular season games in London each season as part of NFL London Games, allowing the league to test solutions to some of the challenges facing a hypothetical London franchise."

So 2007 marked the start of the NFL playing regular season games at the newly rebuilt Wembley Stadium in London as part of the International Series. Prior to that, there were some preseason games played in London, but 2007 was the first year for regular season NFL games in the city.
Reference Answer: since 2007
Correct: Yes
——
Question: where is wind power used most in the united states
Model Answer: Iowa is the state that generates the highest percentage of energy from wind power, with more than 35% of its electric power coming from wind. Alternatively, North Dakota has the highest per capita wind generation.

The state with the most installed wind power capacity is Texas, followed by Iowa. Texas's capacity is around 20,000 MW, almost double that of Iowa's 3,670 MW. The Roscoe Wind Farm in Texas is also the world's largest wind farm at 780 MW.

California has doubled its wind power capacity since 2002 and now gets about 5% of its energy from wind power. This is still lower than the twelve U.S. states that get more than 20% of their energy from wind.
Reference Answer: Texas
Correct: Yes
——
Question: who is the owner of phoenix mall in chennai
Model Answer: The owner of Phoenix Marketcity in Chennai is Phoenix Mills Limited, as mentioned in the document titled "Phoenix Marketcity (Chennai)." This mall was jointly developed by Phoenix Mills Limited and Crest Ventures Ltd.
Reference Answer: Phoenix Mills Limited
Correct: Yes
——
Question: number of degree of freedom for plane mechanism
Model Answer: In mechanics, the degree of freedom (DOF) of a mechanical system is the number of independent parameters that define its configuration. The number of degrees of freedom for a mechanism on a two-dimensional plane is three, which consists of two translations and one rotation.
For a system of multiple bodies, the combined DOF is the sum of the DOFs of the individual bodies, minus the internal constraints on their relative motion. Therefore, a mechanism formed of multiple rigid bodies may have more than three DOFs.
Reference Answer: two
Correct: No
——
Question: who had the best strike rate among batsmen who scored over 200 runs in the world cup
Model Answer: The provided documents do not specify the strike rate of all batsmen who scored over 200 runs in the World Cup. However, among the batsmen mentioned who have scored over 200 runs in the World Cup, Jehan Mubarak has the highest ever strike rate in a T20 World Cup match for Sri Lanka, with a strike rate of 353.84 (Document [6]). But it's important to note that this strike rate is not necessarily reflective of his overall performance in the World Cup, as it could have been achieved in a single match.

For a more accurate comparison of strike rates among batsmen who scored over 200 runs in the World Cup, a comprehensive list of all such batsmen and their respective strike rates would be required.
Reference Answer: Rohit Sharma
Correct: No
——
Question: who played ben stone son on law and order
Model Answer: According to the provided search results, Ben Stone's son Peter Stone was played by Philip Winchester. Some key details:

- Peter Stone is the son of Benjamin "Ben" Stone, the former Executive Assistant District Attorney portrayed by Michael Moriarty on Law & Order.

- Peter first appeared as a character on the Chicago P.D. episode "Justice", which served as a backdoor pilot for the short-lived spin-off Chicago Justice, where he was the lead character played by Philip Winchester.

- After Chicago Justice was canceled, Philip Winchester reprised his role as Peter Stone in Season 19 of Law & Order: Special Victims Unit, where he became the Assistant District Attorney for the Sex Crimes Bureau.

- In several SVU episodes, Peter describes having a strained relationship with his workaholic father Ben Stone, who neglected his family in favor of his prosecutorial duties.

So in summary, Philip Winchester played the role of Peter Stone, the son of the iconic Law & Order character Ben Stone, across the shows Chicago Justice and Law & Order: SVU.
Reference Answer: Michael Moriarty
Correct: No
——
Question: who won so you think you can dance season 13
Model Answer: I'm sorry, I couldn't find any information about the winner of season 13 of So You Think You Can Dance.
Reference Answer: Leon " Kida " Burns
Correct: No
—— Question: when did sweet caroline start at red sox games
Model Answer: The song "Sweet Caroline" started being played at Fenway Park, home of the Boston Red Sox, since at least 1997. However, it has been played consistently in the middle of the eighth inning at every game since 2002.
Reference Answer: at least 1997
Correct: Yes
——
Question: when did end of the road come out
Model Answer: I found multiple results for "End of the Road":

## End of the Road (1944 film)
End of the Road is a 1944 American crime film directed by George Blair and written by Denison Clift, Gertrude Walker and Albert Beich. The film stars Edward Norris, John Abbott, June Storey, Jonathan Hale, Pierre Watkin and Ted Hecht.

## End of the Road (1970 film)
End of the Road is a 1970 American comedy-drama film directed, co-written, and edited by Aram Avakian and adapted from a 1958 novel by John Barth. The film stars Stacy Keach, James Earl Jones and Harris Yulin.

## End of the Road (Boyz II Men song)
End of the Road is a single recorded by American R&B group Boyz II Men for the "Boomerang" soundtrack. It was released on June 30, 1992.

## End of the Road (Jerry Lee Lewis song)
End of the Road is a song by Jerry Lee Lewis. It appeared on the 1976 Sun Records compilation album "The Original Jerry Lee Lewis" which was released by the Sun Record Company.

Which of these were you referring to?
Reference Answer: June 30 , 1992
Correct: Yes
——
Question: who played john connor in the original terminator
Model Answer: John Connor is not physically portrayed by any actor in the original "Terminator" film released in 1984, as he is only mentioned and not seen. The character of John Connor is first referred to in this film as the future leader of the human resistance against the machines, but he does not make an appearance. Reference Answer: Edward Furlong throughout the remainder of the film
Correct: No
——
Question: {QUESTION}
Model Answer: {GEN_ANSWER}
Reference Answer: {GOLD_ANSWER}
Correct:
```

Table 14: KILT NQ GPT-4 optimized Judge Prompt