

Multi-Step Reasoning in Korean and the Emergent Mirage

Guijin Son^{1,2,3} Hyunwoo Ko¹ Dasol Choi²

OneLineAI¹ Yonsei University² MODULABS³
spthsrbls123@yonsei.ac.kr

Abstract

We introduce HRMCR (HAE-RAE Multi-Step Commonsense Reasoning), a benchmark designed to evaluate large language models’ ability to perform multi-step reasoning in culturally specific contexts, focusing on Korean. The questions are automatically generated via templates and algorithms, requiring LLMs to integrate Korean cultural knowledge into sequential reasoning steps. Consistent with prior observations on emergent abilities, our experiments reveal that models trained on fewer than $2 \cdot 10^{25}$ training FLOPs struggle to solve any questions, showing near-zero performance. Beyond this threshold, performance improves sharply. State-of-the-art models (e.g., O1) still score under 50%, underscoring the difficulty of our tasks. Notably, stepwise analysis suggests the observed emergent behavior may stem from compounding errors across multiple steps rather than reflecting a genuinely new capability. We publicly release the benchmark and commit to regularly updating the dataset to prevent contamination.

1 Introduction

Large language models (LLMs) have shown notable success in solving complex reasoning tasks across STEM (Rein et al., 2023) and mathematics (Cobbe et al., 2021), facilitated by methods such as chain-of-thought prompting (Wei et al., 2022b) and inference-time scaling (Brown et al., 2024). However, it is unclear how beneficial these high-level problem-solving abilities—particularly for Olympiad-level math problems—are in real-world, everyday scenarios. Moreover, most existing reasoning benchmarks emphasize universal knowledge (Hendrycks et al., 2021; Fang et al., 2024) that remains constant across different cultures and languages. While these benchmarks (Ko et al., 2025) effectively measure a model’s general reasoning capabilities, they are less suited for evaluating the

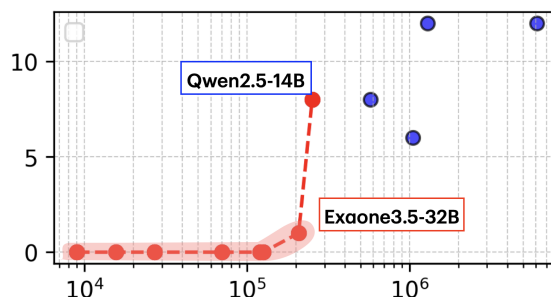


Figure 1: The X-axis represents the training compute scale in ExaFLOPs (10^{18} floating-point operations), calculated as $6 \times \text{\#parameters} \times \text{\#tokens}$ following Kaplan et al. (2020). The Y-axis indicates the performance of the models on HRMCR.

model’s proficiency in language-specific or culturally grounded reasoning.

To address this gap, we introduce HRMCR (HAE-RAE Multi-Step Commonsense Reasoning), a systematically generated benchmark that requires LLMs to integrate Korean cultural knowledge into multi-step reasoning. Our generation algorithm, which uses randomly selected seeds, generates questions and step-by-step solutions. Each of the two subsets in our benchmark comprises 50 questions. We make the *question–answer set* publicly available but withhold the algorithm itself to prevent contamination and overfitting. Instead, we commit to regularly updating the benchmark¹.

As illustrated in Figure 1, models with fewer than $2 \cdot 10^{25}$ training FLOPs struggle to solve any questions on the benchmark. Once training surpasses this threshold, however, scores increase sharply—a phenomenon often described as an emergent ability. Despite this improvement, even state-of-the-art (SOTA) LLMs (e.g., O1, GPT-4o, Claude-3.5-Sonnet) fall below 50% accuracy, underscoring the benchmark’s difficulty. Notably, a stepwise performance analysis indicates that this apparent “emergence” may stem from cumulative

¹Link has been removed for anonymous submission. It will be publicly available.

errors across multiple reasoning steps, rather than a genuinely new capability.

2 Related Works

Korean Benchmarks Most publicly available Korean benchmarks primarily focus on knowledge (Son et al., 2024b; Kim et al., 2024; Son et al., 2023) or safety (Lee et al., 2023a,b), leaving a gap in assessing more advanced reasoning skills. Consequently, the evaluation of Korean LLMs often relies on English benchmarks (Research et al., 2024b,a; Yoo et al., 2024), to evaluate the reasoning capability of LLMs. This highlights the need for Korean-specific reasoning benchmarks that demand both linguistic and cultural competencies.

Multi-Step Reasoning As LLMs continue to enhance their reasoning abilities through improved pre-training (Yang et al., 2024; Lu et al., 2024) and post-training (Wu et al., 2024), many existing benchmarks no longer pose meaningful challenges or offer practical utility. Frontier LLMs (OpenAI, 2024) can now achieve near-expert performance on specialized tasks (Rein et al., 2023; Gao et al., 2024), prompting the creation of even more demanding problem sets (Glazer et al., 2024). However, those tasks often require domain-specific expertise—particularly in STEM—making them less representative of the broader range of reasoning tasks encountered in daily life.

One way to create challenging yet broadly accessible questions is to design multi-step problems that do not require deep expert knowledge. Such tasks can be generated by concatenating simpler subproblems (Hosseini et al., 2024; Son et al., 2024a) or by using template-based algorithms (Sprague et al., 2023), with the latter offering greater diversity. Building on prior work, we develop a fully algorithm-based benchmark that systematically generates multi-step reasoning problems demanding Korean commonsense knowledge. Additionally, we provide an automatic explanation generator, enabling more effective error analysis and evaluation for each step of the reasoning process.

3 HRMCR

In this section, we introduce the HRMCR (HAE-RAE Multi-Step Commonsense Reasoning) benchmark, describing its two subsets (Section 3.1) and key design choices (Section 3.2).

3.1 Subsets

HRMCR comprises two subsets: **Date** and **Zodiac**, each created to require multiple steps of reasoning. We release 50 questions per subset.

Date The Date subset contains concise, two-sentence questions involving Korean holidays and traditional date expressions. To solve these, the model must recall Korean cultural knowledge, perform basic arithmetic, and convert between solar and lunar calendars. Each question requires a five-step solution.

Zodiac The Zodiac subset features longer questions, typically spanning 10–12 lines. These tasks require handling a range of Korean cultural elements—such as the country’s unique age system¹, conversational age expressions, and honorifics. The model must then extract logical inferences from the given premises and perform arithmetic to determine the zodiac sign. Each solution requires seven steps of reasoning at maximum.

3.2 Design Choice

Each subset is generated by a dedicated algorithm comprising roughly 500 lines of code, and each algorithm includes a built-in solution generator that derives a gold-standard solution step-by-step. This built-in solver is a unique feature of our benchmark, as it facilitates comprehensive error analysis at each reasoning stage.

Benchmark contamination has become a pressing issue (Xu et al., 2024; Zhang et al., 2024), and one common approach to address it is using private test sets (Chollet, 2019). However, private benchmarks have drawbacks: over the long term, repeated attempts can lead to overfitting, with models effectively using the test scores as a training signal (Park et al., 2024). Additionally, keeping questions private hinders error analysis and limits opportunities for diagnosing and improving model performance. To balance these concerns, we publicly release our test set but keep the exact generation code confidential. If contamination arises, we can easily regenerate a new question–answer set by varying random seeds. By withholding the generation algorithm, we ensure that every newly released set remains unseen, making the benchmark more robust over time.

Algorithm 1 illustrates the pseudo-code used to create the Date questions, and the algorithm for the

¹Korea calculates age differently from Western countries.

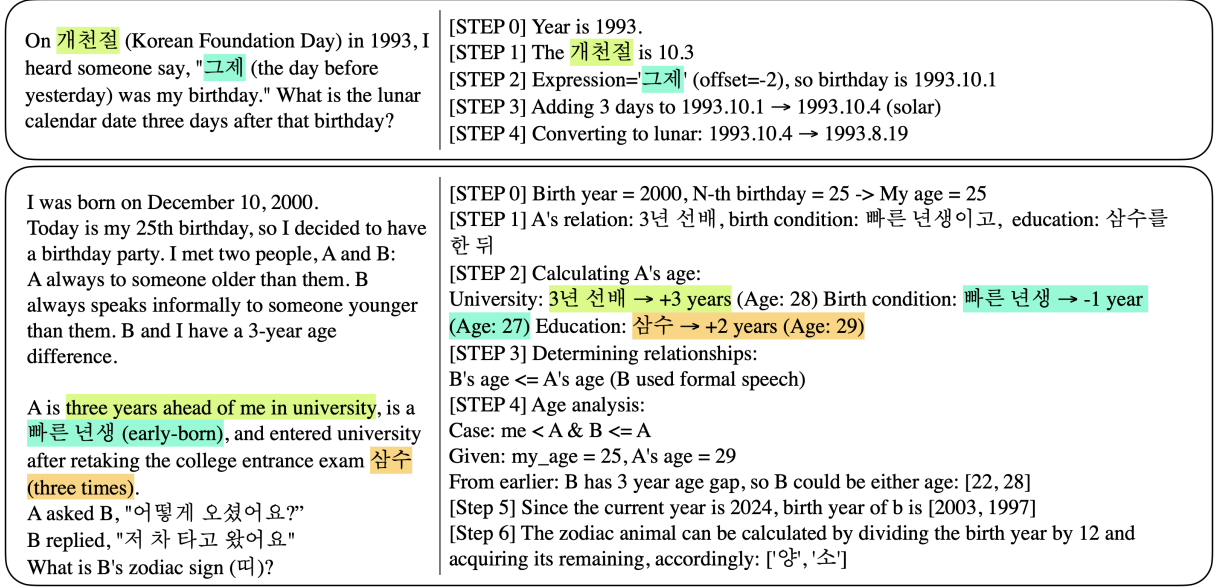


Figure 2: Example of generated questions in the HRMCR benchmark. The figure showcases generated questions (left) alongside their automatically generated solutions (right). The top panel represents the "date" subset, while the bottom corresponds to the "zodiac" subset. Questions are translated into Korean to enhance accessibility.

Algorithm 1 Question Generator for Date Subset

Require: Database of cultural events with calendar types and dates

```

1: function GENERATEQUESTION
2:   // Step 1: Select base components
3:   year, event ← Random(valid_year_range), RandomSelect(cultural_events)
4:   cal_type ← event.calendar_type
5:   // Step 2: Select expressions
6:   date_expr, target_cal ← RandomSelect(date_expressions), RandomSelect(calendar_types)
7:   // Step 3: Generate question
8:   question ← Template( year, event.name, date_expr, target_cal)
   return question
9: end function

```

Zodiac subset is provided in Appendix A. Figure 2 and 5 show examples of generated questions and their gold solutions.

4 Experimental Setup

In this section, we describe how responses were generated (Section 4.1) and evaluated (Section 4.2).

4.1 Response Generation

We evaluate a total of 20 LLMs, including proprietary models such as GPT-4o, GPT-4o-Mini (Hurst et al., 2024), O1, O1-Mini (Jaech et al., 2024), and Claude-3.5-Sonnet (Anthropic, 2024), as well as open models like Qwen2.5 (Yang et al., 2024), Llama3 (Dubey et al., 2024), Exaone3.5 (Research et al., 2024a), and DeepSeek3 (DeepSeek-AI et al., 2024). All models are evaluated in a greedy setting. For models up to 32B parameters, we run inference

on a local GPU server; larger models are accessed via the OpenRouter API.¹ For additional details on the evaluated models, see Appendix A.

4.2 Response Evaluation

For evaluation, we use GPT-4o as an LLM-as-a-Judge (Zheng et al., 2023). The judge takes the question, the model-generated response, and the gold step-by-step solution. As shown in Figure 4 (Appendix A), the judge first provides a brief comparison with the gold solution and then determines whether the model’s response is correct. If it is incorrect, the judge identifies the specific step at which the error occurred. All evaluations use greedy decoding.

¹<https://openrouter.ai/>

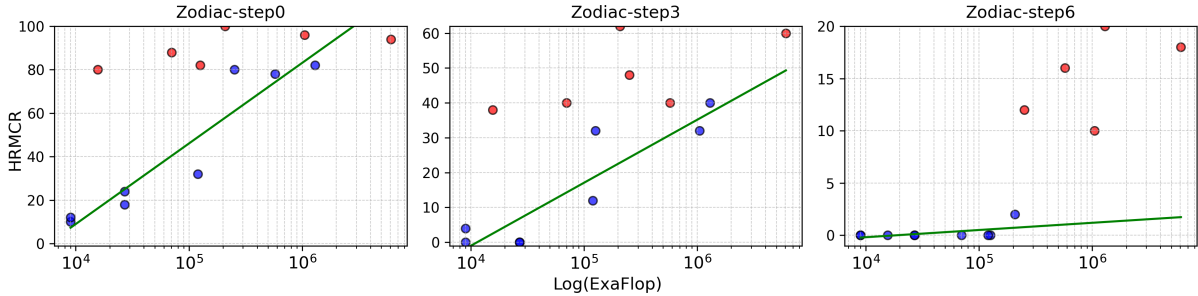


Figure 3: Breakdown of performance results for selected steps in the Zodiac subset. The green line represents the regression line, the blue points are instances used for fitting the regression, and the red points represent the test set. Only at the last step, the regression fails. For the entire results, see Appendix B.

5 Evaluation Results

Models	Date	Zodiac	Av.
o1	34	56	45
GPT-4o	28	32	30
DeepSeek-V3	32	14	23
Claude-3.5-Sonnet	34	8	21
Qwen2.5-72B	4	20	12
Llama3.1-405B	6	18	12
EXAONE3.5-32B	0	2	1

Table 1: Evaluation results on HRMCR. We only display the performance of top-performing models per model family. The best-performing model is highlighted in **bold**.

Table 1 presents the evaluation results, from which we derive three key observations. First, the HRMCR benchmark is highly challenging: leading models such as GPT-4o, DeepSeek-V3, and Claude-3.5-Sonnet all score under 30%. This is particularly noteworthy given that the benchmark is built on fixed, deterministic rules rather than specialized domain knowledge. Second, OpenAI’s latest reasoning-oriented LLM, O1, achieves an average score of 45, substantially outperforming earlier models. This suggests that inference-time scaling can generalize effectively to previously unseen domains. Finally, EXAONE3.5-32B, despite its size, shows near-zero performance on the benchmark. This indicates that solving HRMCR requires not just model scale but also advanced training strategies and sufficient computational resources.

6 The Emergent Mirage

Emergent at First Sight Emergent capability refers to abilities absent in smaller models but present in larger ones, making them unpredictable based solely on the performance of smaller models (Wei et al., 2022a). In Figure 1, we plot the per-

formance of 19 models, with log compute on the X-axis and average performance on HRMCR on the Y-axis. Models exhibit near-zero performance until reaching $2 \cdot 10^{25}$ training FLOPs, followed by a sudden upsurge between Exaone3.5-32B and Qwen2.5-14B. This indicates that performance is driven primarily by training compute rather than model size: Qwen2.5-14B outperforms EXAONE3.5-32B by training on three times more tokens.

Alternative Interpretation Wei et al. (2022a) suggests that the sudden "emergence" of abilities may be attributed to multi-step reasoning. To investigate this hypothesis in our benchmark, we analyze the performance at each intermediate step. In Figure 3, we plot step-wise accuracy and test for emergent behavior using linear regression. We train a regression model on smaller models (excluding the top five performers) to predict performance at each step. Surprisingly, while this regression accurately predicts the performance of larger models at all intermediate steps, it fails only at the final step. This pattern holds for both Date and Zodiac subsets, despite their final steps involving different types of operations. Given that the final step in the Zodiac subset involves basic arithmetic (simple division), which is not typically considered an emergent capability, we propose an alternative interpretation: rather than true emergence, this pattern may result from error accumulation across steps. We hypothesize that further decomposing tasks into smaller steps would reveal purely linear performance trends, challenging the notion of emergence in this context.

7 Conclusion

We introduced HRMCR, a Korean multi-step reasoning benchmark that combines cultural knowledge with systematic generation. Our analysis re-

vealed that seemingly emergent capabilities in LLMs may be artifacts of accumulated errors, prompting a reconsideration of how we evaluate model capabilities.

Acknowledgments

This research was supported by Brian Impact Foundation, a non-profit organization dedicated to the advancement of science and technology for all.

References

- AI Anthropic. 2024. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 3.
- Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.
- François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojuan Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shanyang Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjuan Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. *Deepseek-v3 technical report*. *Preprint*, arXiv:2412.19437.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. 2024. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. *arXiv preprint arXiv:2406.18321*.
- Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.
- Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- Arian Hosseini, Alessandro Sordani, Daniel Toyama, Aaron Courville, and Rishabh Agarwal. 2024. Not all llm reasoners are created equal. *arXiv preprint arXiv:2410.01748*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar,

- Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. Click: A benchmark dataset of cultural and linguistic intelligence in korean. *arXiv preprint arXiv:2403.06412*.
- Hyunwoo Ko, Guijin Son, and Dasol Choi. 2025. Understand, solve and translate: Bridging the multilingual mathematical reasoning gap. *arXiv preprint arXiv:2501.02448*.
- Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Meeyoung Cha, Yejin Choi, Byoung Pil Kim, Gunhee Kim, Eun-Ju Lee, Yong Lim, et al. 2023a. Square: A large-scale dataset of sensitive questions and acceptable responses created through human-machine collaboration. *arXiv preprint arXiv:2305.17696*.
- Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-Woo Ha. 2023b. Kosbi: A dataset for mitigating social bias risks towards safer large language model application. *arXiv preprint arXiv:2305.17701*.
- Zimu Lu, Aojun Zhou, Ke Wang, Houxing Ren, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. 2024. Mathcoder2: Better math reasoning from continued pretraining on model-translated mathematical code. *arXiv preprint arXiv:2410.08196*.
- OpenAI. 2024. [Learning to reason with llms](#).
- Chanjun Park, Hyeonwoo Kim, Dahyun Kim, Seonghwan Cho, Sanghoon Kim, Sukyung Lee, Yungi Kim, and Hwalsuk Lee. 2024. Open ko-llm leaderboard: Evaluating large language models in korean with ko-h5 benchmark. *arXiv preprint arXiv:2405.20574*.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.
- LG Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, et al. 2024a. Exaone 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:2412.04862*.
- LG Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Stanley Jungkyu Choi, Yemuk Choi, Seokhee Hong, Yeonjung Hong, Junwon Hwang, Hyojin Jeon, et al. 2024b. Exaone 3.0 7.8 b instruction tuned language model. *arXiv preprint arXiv:2408.03541*.
- Guijin Son, Sangwon Baek, Sangdae Nam, Ilgyun Jeong, and Seungone Kim. 2024a. Multi-task inference: Can large language models follow multiple instructions at once? *arXiv preprint arXiv:2402.11597*.
- Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024b. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*.
- Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jaechol Lee, Je Won Yeom, Jihyu Jung, Jung Woo Kim, and Songseong Kim. 2023. Hae-rae bench: Evaluation of korean knowledge in language models. *arXiv preprint arXiv:2309.02706*.
- Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2023. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*.
- Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. 2024. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Kang Min Yoo, Jaeyeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. 2024. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*.
- Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, et al. 2024. A careful examination of large language model performance on grade school arithmetic. *arXiv preprint arXiv:2405.00332*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,

Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Evaluated Models

Llama-3 (Dubey et al., 2024). The Llama-3 series, spanning from Llama-3 to Llama-3.3, comprises large language models ranging from 1 to 405 billion parameters developed by Meta. While these multilingual models are pretrained on datasets comprising 9 to 15 trillion tokens across multiple languages, they do not officially support Korean. The suite also provides instruction-tuned models.

Qwen2.5 (Yang et al., 2024). Qwen2.5 is a suite of multilingual language models ranging from 0.5 to 72 billion parameters developed by Alibaba. Each model within the series is pretrained on a dataset of 18 trillion multilingual tokens, including Korean. Furthermore, the Qwen2.5 series officially support Korean and includes instruction-tuned variants for instruction-following.

DeepSeek-V3 (DeepSeek-AI et al., 2024). DeepSeek-V3 is a Mixture-of-Experts (MoE) language model with 671 billion parameters, utilizing 37 billion active parameters per token. It is trained on a dataset of 14.8 trillion multilingual tokens, making it robust across diverse languages and contexts. Additionally, they introduce the instruction-tuned version of the model.

EXAONE-3.5 (Research et al., 2024a). EXAONE 3.5 is a suite of multilingual language models with sizes of 2.4B, 7.8B, and 32B parameters developed by LG AI Research. These models are pretrained on datasets comprising up to 9 trillion tokens, evenly balanced between Korean and English, to ensure strong bilingual capabilities. Only the instruction-tuned versions of the models were made publicly available.

GPT-4o & o1 (Hurst et al., 2024; Jaech et al., 2024). GPT-4o, an advanced version of GPT-4, accepts multimodal inputs—including text, audio, image, and video—and demonstrate significant improvements on text in non-English languages. The o1 is a model designed to perform high-quality reasoning on complex tasks that require extensive thought processes, leveraging inference-time scaling through more elaborate reasoning steps and reflection. Both models are developed by OpenAI, and the latest versions of each were utilized in the experiments.

Claude-3.5 (Anthropic, 2024). Claude-3.5 series is the next generation of the Claude

Algorithm 2 Question Generator for Zodiac Subset

Require: Database of relationships, speech levels, and age modifiers

Ensure: Question about age relationships and zodiac sign

```
1: function GENERATEQUESTION
2:   // Step 1: Generate base timeline
3:   birth_year, current_year  $\leftarrow$  Random(valid_range), Random(valid_range)
4:   my_age  $\leftarrow$  CalculateKoreanAge(birth_year, current_year)
5:   // Step 2: Generate person A's profile
6:   relationship  $\leftarrow$  RandomSelect(university_relationships)
7:   modifiers  $\leftarrow$  RandomSelect(age_modifiers)
8:   a_age  $\leftarrow$  CalculateAge(my_age, relationship, modifiers)
9:   // Step 3: Generate conversation
10:  speech_level_a, speech_level_b  $\leftarrow$  RandomSelect(speech_levels), RandomSelect(speech_levels)
11:  age_relationship  $\leftarrow$  DetermineRelationship(speech_level_a, speech_level_b)
12:  // Step 4: Generate question text
13:  question  $\leftarrow$  Template(my_age, relationship, modifiers, conversation)
   return question
14: end function
```

3 series with significant improvements in commonsense and STEM reasoning benchmarks developed by Anthropic. We utilize the claude-3.5-sonnet-20241022 for the experiments.

B Details in Evaluation

In this section, we provide samples of the prompts used for evaluation, responses generated by GPT-4o as LLM-as-a-Judge, and additional evaluation results.

B.1 Evaluation Prompt

We use GPT-4o as an LLM-as-a-Judge for evaluation; in Figure 4, we provide the prompt used for evaluation. The evaluation is done in greedy decoding settings.

B.2 Sample Responses

In Figure 5, we provide sample responses by models, and judgements generated by GPT-4o.

B.3 Additional Results

In this section, we present additional results. For the performance of all 20 models broken down by steps, see Table 2 and 3. In Figures 7 and 6, we provide continue from Section 6 and provide regression results for all steps of both subsets. Our findings remain consistent with the additional results.

System Prompt:
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the provided question. You will be given a question, a gold step-by-step answer, and a response from an AI assistant. Review the response by the AI assistant. Compare it with the gold step-by-step answer.

Begin your evaluation by providing a comparison with the gold answer. Be as objective as possible. After providing your explanation, return whether the model has reached the correct answer, and if not specify which step it has first failed at. For example:

if correct -> correct: [[true]] step: [[-1]]
if wrong -> correct: [[false]] step: [[3]]

User Prompt:
Question :
{question}

Gold Answer :
{gold}

Model Response :
{response}

Figure 4: Prompt used for judging responses.

Models	Step 0	Step 1	Step 2	Step 3	Final
Exaone3.5-2.4B	94	38	20	6	0
Exaone3.5-7.8B	86	56	42	14	0
Exaone3.5-32B	100	74	52	16	0
Qwen2.5-0.5B	56	30	14	6	0
Qwen2.5-1.5B	42	12	4	0	0
Qwen2.5-7B	100	72	40	22	0
Qwen2.5-14B	100	82	50	30	4
Qwen2.5-32B	100	76	54	28	0
Qwen2.5-72B	100	74	50	30	4
Llama3.2-1B	48	26	14	2	0
Llama3.2-3B	84	48	30	6	0
Llama3.1-8B	100	70	38	8	0
Llama-3.3-70B	100	78	48	14	2
Llama3.1-405B	100	72	48	24	6
DeepSeek3	100	70	54	40	32
Claude-3.5-Sonnet	100	70	60	48	34
GPT-4o-mini	100	68	44	32	10
GPT-4o	100	76	54	40	28
o1-mini	100	76	46	32	0
o1	100	76	64	60	34

Table 2: Performance of all models on the Date subset, broken down per step.

Question:
 On 새해 첫날 (New Year's Day) in 1999, I heard someone say, "이튿날" (the day after tomorrow) was my birthday. What is the lunar calendar date 57 days after that birthday?

Gold Solution:
 STEP 0: Year is 1999.
 STEP 1: The 새해 첫날 is 1.1
 STEP 2: Expression='이튿날' (offset=+1), so birthday is 1999.1.2
 STEP 3: Adding 57 days to 1999.1.2 → 1999.2.28 (solar)
 STEP 4: Converting to lunar: 1999.2.28 → 1999.1.13

Response by O1:
 1999년 음력 설날(새해 첫날)은 1월 1일로, 그 다음 날인 1월 2일이 생일입니다. 이 생일로부터 57일 후의 음력 날짜를 계산해 보겠습니다.

1. 1월 2일부터 1월 말까지:
 - 1월은 30일로 구성되어 있으므로, 1월 2일부터 1월 30일까지는 총 29일입니다.

2. 2월:
 - 1999년 음력 2월은 29일입니다.
 - 57일에서 1월에 사용된 29일을 제외하면, 28일이 남습니다.
 - 따라서 2월 1일부터 2월 28일까지가 됩니다.

결론적으로, 1월 2일로부터 57일 후의 음력 날짜는 2월 28일입니다.

답: 음력 2월 28일

LLM-as-a-Judge:
 The AI assistant's response does not align with the gold step-by-step answer. Here's a comparison:

1. Year Identification:
 - Both the gold answer and the AI response correctly identify the year as 1999.

2. New Year's Day Identification:
 - The gold answer identifies the solar New Year's Day as January 1st, 1999, and the birthday as January 2nd, 1999.
 - The AI response incorrectly assumes the lunar New Year's Day as January 1st, 1999, and the birthday as January 2nd, 1999, without clarifying the calendar system.

3. Calculation of 57 Days After Birthday:
 - The gold answer correctly calculates 57 days after January 2nd, 1999, in the solar calendar, resulting in February 28th, 1999.
 - The AI response incorrectly calculates the days using the lunar calendar, leading to a different conclusion.

4. Conversion to Lunar Date:
 - The gold answer converts February 28th, 1999, from the solar calendar to the lunar calendar, resulting in the lunar date of January 13th, 1999.
 - The AI response does not perform this conversion and instead provides a lunar date of February 28th, which is incorrect.

The AI assistant fails to correctly interpret the calendar system and does not perform the necessary conversion from the solar to the lunar calendar. The first failure occurs at step 3, where the AI incorrectly calculates the 57 days using the lunar calendar instead of the solar calendar.

Correct: [[false]] Step: [[3]]

Figure 5: A full example including a sample question, gold solution, model response (by o1) and judgment (by GPT-4o).

Models	Step 0	Step 1	Step 2	Step 3	Step 4	Step 5	Final
Exaone3.5-2.4B	80	76	54	38	34	4	0
Exaone3.5-7.8B	88	84	52	40	34	8	0
Exaone3.5-32B	100	98	66	62	40	22	2
Qwen2.5-0.5B	10	4	0	0	0	0	0
Qwen2.5-1.5B	24	16	2	0	0	0	0
Qwen2.5-7B	82	68	36	32	26	10	0
Qwen2.5-14B	80	80	58	48	42	28	12
Qwen2.5-32B	78	78	46	40	38	28	16
Qwen2.5-72B	82	80	40	40	32	28	20
Llama3.2-1B	12	12	12	4	0	0	0
Llama3.2-3B	18	8	0	0	0	0	0
Llama3.1-8B	32	22	14	12	4	2	0
Llama-3.3-70B	96	92	52	32	30	20	10
Llama3.1-405B	94	90	72	60	54	32	18
DeepSeek3	94	94	70	60	48	20	14
Claude-3.5-Sonnet	32	30	26	14	8	8	8
GPT-4o-mini	96	90	72	64	52	22	20
GPT-4o	88	84	54	50	42	32	32
o1-mini	92	90	70	64	48	44	44
o1	98	98	72	68	58	56	56

Table 3: Performance of all models on the Zodiac subset, broken down per step.

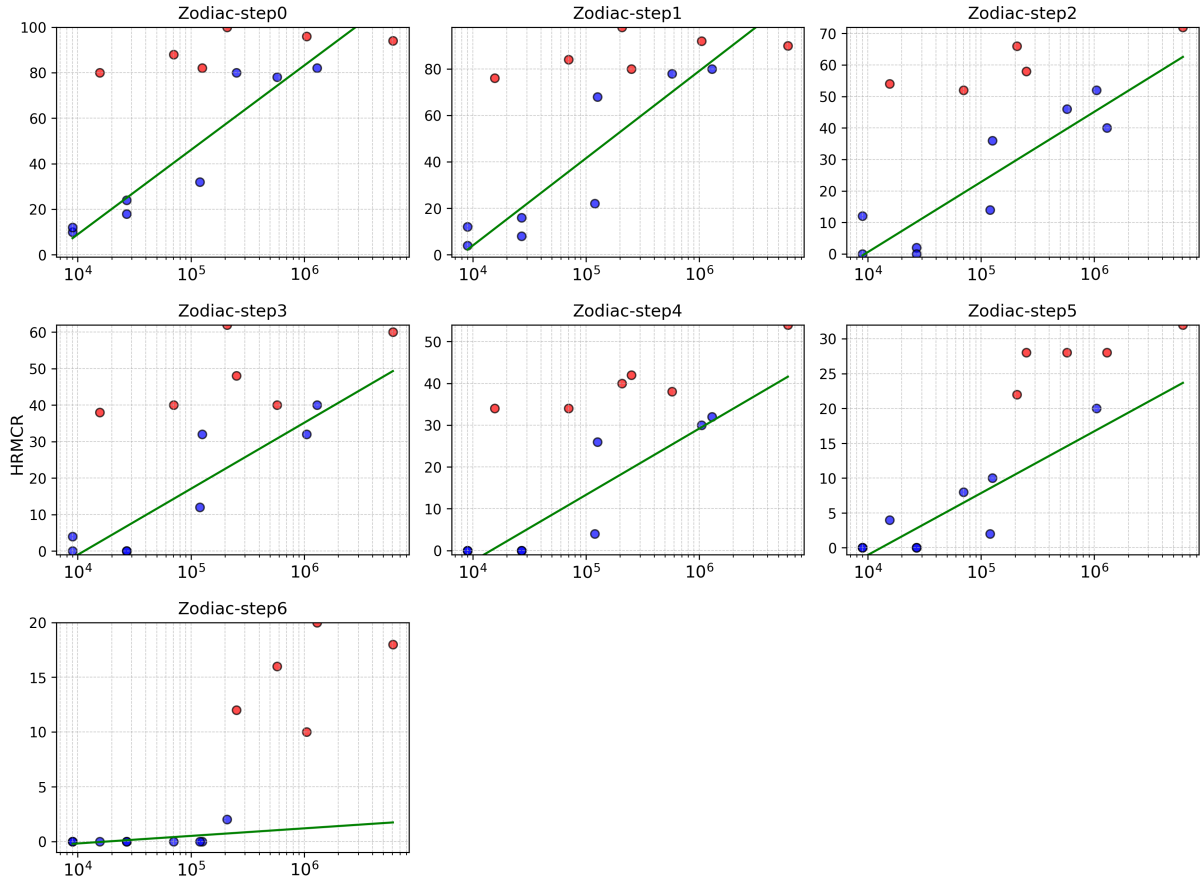


Figure 6: Breakdown of performance results for all steps in the Zodiac subset. The green line represents the regression line, the blue points are instances used for fitting the linear regression, and the red points represent the test set. Only at the final step the regression fails.

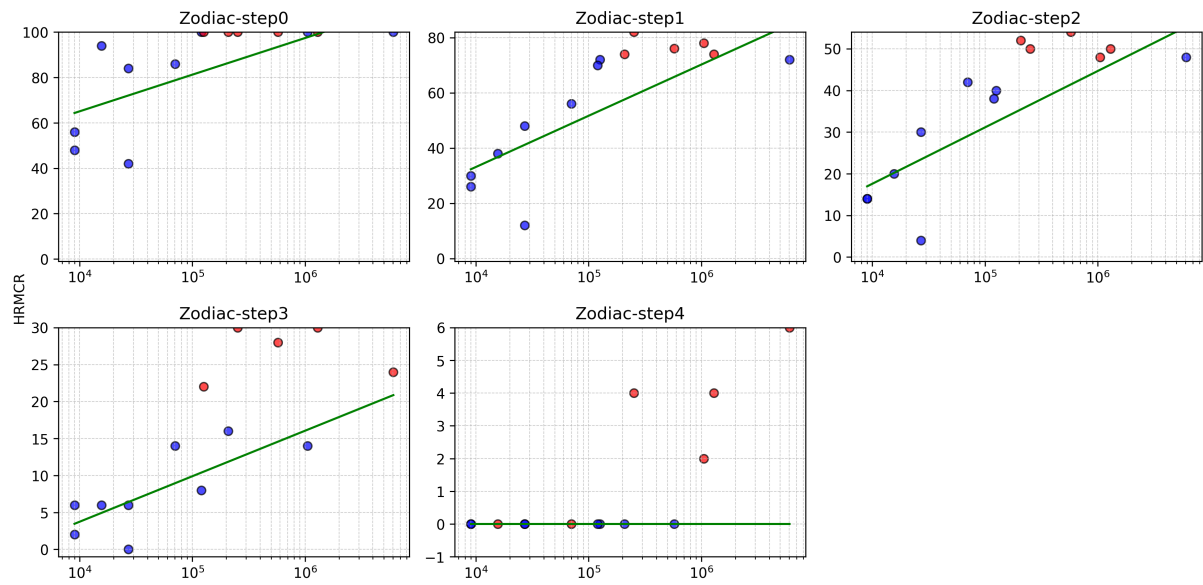


Figure 7: Breakdown of performance results for all steps in the Date subset. The green line represents the regression line, the blue points are instances used for fitting the linear regression, and the red points represent the test set. Only at the final step the regression fails.