S2L-RM: Short-to-Long Reward Modeling

Changyu Chen[*12], **Zichen Liu**[13], **Haonan Wang**[13], **Chao Du**[†1], **Tianyu Pang**[1], **Qian Liu**[1], **Arunesh Sinha**[4], **Pradeep Varakantham**[2], **Min Lin**[1]
[1]Sea AI Lab, Singapore [2]Singapore Management University
[3]National University of Singapore [4]Rutgers University
{chency,liuzc,wanghn,duchao,tianyupang,liuqian,linmin}@sea.com;
arunesh.sinha@rutgers.edu; pradeepv@smu.edu.sg

Abstract

Preference tuning has been effective in aligning language models with human values, often relying on reward models to annotate preferences for generated responses. However, extending this stage to long context language models requires reward models capable of accurately evaluating responses of long context tasks — a challenge that current models struggle to address despite their expanded context windows. We introduce S2L-RM, an approach that leverages short context reward models to assess the responses of long context tasks. Our method employs a factual verifier to select responses within a trust region relative to a reference response. These responses are then evaluated using any short context reward model, with input limited to a short query, the reference response, and the model-generated response. Our preliminary experiments demonstrate that our approach can accurately provide preference annotations in long-context scenarios.

1 Introduction

The capability of processing long-form inputs holds the key to empower large language models (LLMs) to excel challenging tasks such as book summarization, many-shot in-context learning, and designing complex LLM agents (Bai et al., 2023; Liu et al., 2023; Bai et al., 2024a). Existing work on long-context LLMs has primarily focused on the stage of pre-training (Chen et al., 2023; Hu et al., 2024) or supervised fine-tuning (SFT) (Bai et al., 2024a), leaving alignment from human preferences under-explored. However, as LLMs are getting increasingly powerful, aligning them with human values is of critical importance towards harmfulless superintelligence.

Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020) is a promising approach for LLM alignment and has been successfully applied to various state-of-the-art systems (Bai et al., 2022; Ouyang et al., 2022; Touvron et al., 2023). Yet, long-context RLHF poses a unique challenge that acquiring abundant and high-quality human feedback for long-form prompts is extremely hard. To circumvent this difficulty, ChatGLM (2024) attempts using existing long-context LLMs such as GLM4-128K (GLM et al., 2024) to serve as a judge (Bubeck et al., 2023; Zheng et al., 2023) and rank model completions given long-form inputs, but concludes with negative results. To our knowledge, there lacks an effective method to build a reward model (RM) for long-context RLHF.

In this paper, we explore the potential of leveraging **strong short-context RMs** (Lambert et al., 2024) and readily available **long-context SFT datasets** (Bai et al., 2024a,b) to construct *preference data for long-context RLHF*. This work focuses on the long-context understanding scenario, where the input is typically in the form of $\boldsymbol{x} = (\boldsymbol{c}, \boldsymbol{q})$, with $\boldsymbol{c}$ being a long-form context and $\boldsymbol{q}$ being a short-form

[*]The project was done during Changyu Chen's internship at Sea AI Lab.
[†]Correspondence to Chao Du.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

query. LLMs are tasked to understand the rich context and generate a response $\boldsymbol{y}$ for the query $\boldsymbol{q}$. In order to conduct RLHF, one needs to build an RM that scores individual answers as $r(\boldsymbol{c}, \boldsymbol{q}, \boldsymbol{a})$ in the absence of expert labeled preference data.

To approach this problem, we first assume that labels from the SFT dataset (i.e., reference responses $\boldsymbol{y}_{\text{ref}}$) capture necessary factual information for answering the corresponding long-form inputs. Based on this assumption, we first employ a factual verifier to construct a trust region of completions centered around $\boldsymbol{y}_{\text{ref}}$ for each input $\boldsymbol{x}$. We then evaluate candidate completions within the trust region using any short-context reward model with input ignoring the contexts but incorporating the reference responses. Intuitively, if our assumption holds, $\boldsymbol{y}_{\text{ref}}$ can safely substitute $\boldsymbol{c}$ with minimal information loss, hence our reward modeling can effectively generate trustworthy preference data. We refer to our approach as Short-to-Long Reward Modeling (`S2L-RM`).

To evaluate the effectiveness of `S2L-RM`, we first verify our assumption in a short-context scenario where $r(\boldsymbol{c}, \boldsymbol{q}, \boldsymbol{a})$ can be exactly computed with a "golden" RM. Based on positive results on assumption verification, we further test our approach in the RLHF setting by employing Best-of-N (Nakano et al., 2021; Touvron et al., 2023) as the policy optimization method. The encouraging results obtained validate our approach and motivate further study on preference tuning long-context LLMs with `S2L-RM`.

2 Modeling Rewards for Long Context Language Models

In this section, we first introduce the approach to training reward models for capturing human preferences in scenario with a large amount of long context preference data. We then explore how to extend reward signals from short-context reward models to long-context short prompts.

2.1 Long Context Reward Models in The Ideal Scenario

In the ideal scenario, where substantial long-context preference data is available, the training process for a long-context reward model mirrors that of a short-context reward model. Given pairwise preference data, each prompt $\boldsymbol{x}$ is associated with two possible responses, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. The preference label $o(\boldsymbol{y}_1 \succ \boldsymbol{y}_2 | \boldsymbol{x}) \in \{0, 1\}$ indicates whether $\boldsymbol{y}_1$ is preferred over $\boldsymbol{y}_2$. The preferred response is denoted as $\boldsymbol{y}^+$, while the other is denoted as $\boldsymbol{y}^+$. A common assumption is that the ground-truth human preferences follow the Bradley-Terry model (**?**). Based on this assumption, we can train a parameterized reward model $r_\phi(\boldsymbol{x}, \boldsymbol{y})$ using maximum likelihood:

$$\mathcal{L}_R\left(r_\phi, \mathcal{D}\right) = -\mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}^+, \boldsymbol{y}^-) \sim \mathcal{D}}\left[\log \sigma\left(r_\phi\left(\boldsymbol{x}, \boldsymbol{y}^+\right) - r_\phi\left(\boldsymbol{x}, \boldsymbol{y}^-\right)\right)\right], \tag{1}$$

where σ is the logistic function.

However, collecting long-context preference data is inherently challenging, as evaluating long-context responses is difficult even for humans . To circumvent this, GLM et al. (2024) proposed using a short-context reward model that only considers the query and response while ignoring the long context. This approach, however, can lead to poor evaluation outcomes because the reward model cannot assess the factuality of a response without considering the full context. We illustrate this issue with an example in Section 2.2.

2.2 S2L-RM: Short-to-Long Reward Models

Our approach utilizes a short context reward model. Besides the query and responses, we provide an additional reference response to help for the reward model assessment. The model consists of two components, a factual verifier, $\mathcal{V}$, and a short context reward model, r_{S}.

The factual verifier $\mathcal{V} : \mathcal{Q} \times \mathcal{Y} \times \mathcal{Y} \mapsto \{0, 1\}$ checks the factuality consistency of the response $\boldsymbol{y}$ against the reference response $\boldsymbol{y}_{\text{ref}}$. If $\boldsymbol{y}$ contradicts $\boldsymbol{y}_{\text{ref}}$, $\mathcal{V}$ outputs 0; otherwise, it outputs 1. The short context reward model $r_{\text{S}} : \mathcal{Q} \times \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ then evaluates the quality of $\boldsymbol{y}$ relative to $\boldsymbol{y}_{\text{ref}}$.

We assume that the reference-based reward score $r_{\text{S}}(\boldsymbol{q}, \boldsymbol{y}_{\text{ref}}, \boldsymbol{y})$ aligns well with the long-context reward model r_{L} when $\boldsymbol{y}$ passes factual verification by $\mathcal{V}$. Formally, this can be described as

$$|r_{\mathrm{L}}(\boldsymbol{x}, \boldsymbol{y}_i) - r_{\mathrm{S}}(\boldsymbol{q}, \boldsymbol{y}_{\mathrm{ref}}, \boldsymbol{y}_i)| < \epsilon, \tag{2}$$

for responses $\boldsymbol{y}_i \in \{\boldsymbol{y}_i | \mathcal{V}(\boldsymbol{q}, \boldsymbol{y}_{\mathrm{ref}}, \boldsymbol{y}_i) = 1\}$, where ϵ is a small number. We name the space $\{\boldsymbol{y}_i | \mathcal{V}(\boldsymbol{q}, \boldsymbol{y}_{\mathrm{ref}}, \boldsymbol{y}_i) = 1\}$ the response trust region. If this assumption holds, we can effectively evaluate the responses given the long-form prompts $\boldsymbol{x}$ using a short context RM.

To validate the assumption, we simulate a scenario where the reward model is able to evaluate the responses with the long-form prompts. Specifically, we limit the sequence length including the context to less than 3096. Then we compare three approaches: 1) $r_{\mathrm{S}}(\boldsymbol{q}, \boldsymbol{y})$, only queries and responses are given, discarding context information; 2) `+ref`, where a reference response is given and evaluate the responses by $r_{\mathrm{S}}(\boldsymbol{q}, \boldsymbol{y}_{\mathrm{ref}}, \boldsymbol{y})$; 3) `+ref, +trust`, our approach, all approaches using the same reward model ArmoRM-Llama3-8B-v0.1. We generate N responses for each prompt in a prompt set, then we select the best and the worst ones from N responses using a strong RM (Skywork-Reward-Gemma-2-27B[1]), $N = \{2, 3, ..., 8\}$ in this experiment, which is considered to be the golden ranking as the strong RM takes as input the full prompt $\boldsymbol{x}$ and response $\boldsymbol{y}$. Then, we compute the Kendall correlation coefficient (Kendall, 1948) between the ranking given by different approaches and the golden ranking.

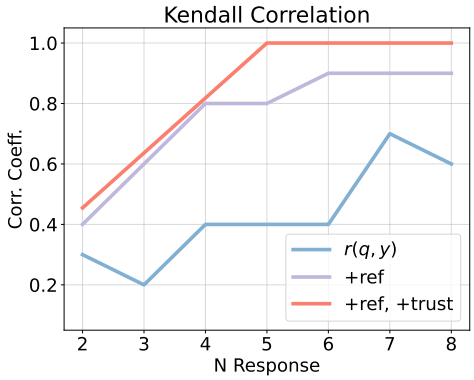

Figure 1: Kendall correlation between the golden ranking and the ranking given by various approaches.

As shown in Figure 1, the vanilla approach, $r(\boldsymbol{q}, \boldsymbol{y})$, has the lowest correlation with the golden ranking, while introducing a reference response (`+ref`) helps to improve the correlation and our approach (`+ref, +trust`) shows the highest correlation with the golden ranking. Moreover, we have observed an increasing trend of the correlation with the increase of N. It makes sense as the best response and the worst response might have only subtle differences with small N, which can cause high disagreement between the golden ranking and the others. This simulated experiment provides some initial evidence to support our assumption. Next, we will further verify our approach on the representative long context benchmark.

3 Empirical Studies

In this section, we empirically evaluate S2L-RM. Our findings highlight several key points: (1) S2L-RM effectively ranks and selects high-quality responses; (2) Even the lowest-ranked responses by S2L-RM exhibit higher quality compared to those of the baseline approach, giving RLHF practitioners greater flexibility in selecting dispreferred responses ($\boldsymbol{y}^-$) for constructing preference datasets; (3) We provide illustrative examples of responses selected by different methods.

3.1 Experiment setup

Models. We utilize `ArmoRM-Llama3-8B-v0.1`[2] as the reward model r_{S} to generate the scalar score. For factual verifier $\mathcal{V}$, we use `GPT-4o-mini` (Wu et al., 2024) to ensure the accuracy in filtering out the data within the trust region. We use two long context language models, `glm-4-9b-chat-1m`[3] and `Meta-Llama-3.1-8B-Instruct`[4], to propose responses for long context tasks.

Baseline. We compare with the approach that discards the long-form context and takes only the query and response as the input, $r(\boldsymbol{q}, \boldsymbol{y})$. This approach is adopted by ChatGLM (2024).

[1] https://huggingface.co/Skywork/Skywork-Reward-Gemma-2-27B
[2] https://huggingface.co/RLHFlow/ArmoRM-Llama3-8B-v0.1
[3] https://huggingface.co/THUDM/glm-4-9b-chat-1m
[4] https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

Evaluation Protocol. We evaluate our method on LongBench-Chat (Bai et al., 2024a), a benchmark that is dedicated to evaluating the long context alignment performance of language models. In LongBench-Chat, 30 question data are for best mimicking real user queries, while the remaining 20 questions are selected from long dependency QA tasks in the LooGLE dataset (Li et al., 2023). In experiments, we sample N responses, rank them by different approaches, select the top 1 and the bottom 1 responses, and evaluate the quality of the responses on LongBench-Chat.

3.2 S2L-RM Effectively Ranks Model Responses

From the experiment results (see Figure 2), if looking at scores of the best responses, we observe that S2L-RM always shows a higher score than the baseline approach $r(\boldsymbol{q}, \boldsymbol{y})$ at 3 sample size scales across two base models. This indicates that S2L-RM is more robust toward the model responses compared with the baseline approach when we use a reward model to work with Best-of-N (Nakano et al., 2021; Touvron et al., 2023) as a policy optimization method. Notably, `Meta-Llama-3.1-8B-Instruct` with the baseline approach has encountered reward hacking where the selected "best" response gets a lower score with $N = 8$ compared with $N = 4$.

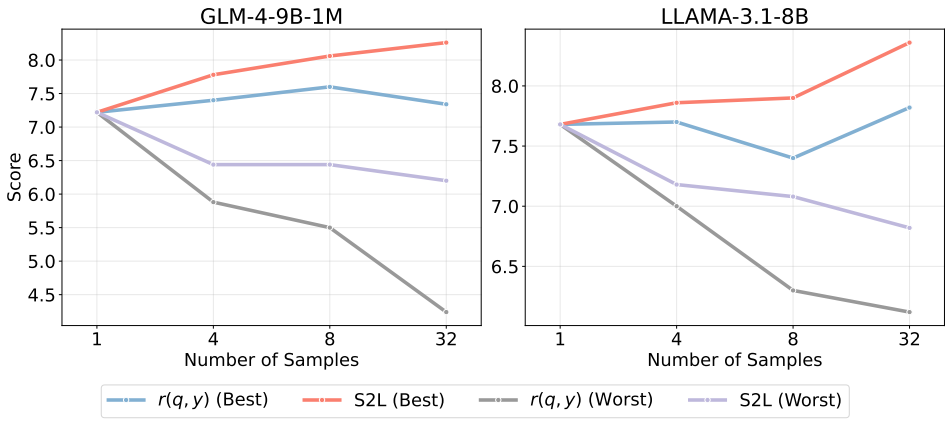

Figure 2: Evaluation results on LongBench-Chat.

If looking at the scores of the worst responses, we observe that the baseline approach can find the responses worse than the one found by S2L-RM. Intuitively speaking, this can be explained by the mechanism that S2L-RM chooses the responses from the trust region, i.e. the factual information of these responses is consistent with the reference response. So even the worst response within the trust region can have higher scores than the one outside the trust region which may include factual errors. The example of responses selected by different approaches is demonstrated in Table 1 in the Appendix.

Based on the properties shown in the experiments, we conclude that different strategies can be utilized to construct the preference dataset. One strategy can be selecting $\boldsymbol{y}^+$ using S2L-RM and $\boldsymbol{y}^-$ using the baseline approach. In this way, the preference could potentially have high diversity but may have the risk of getting $\boldsymbol{y}^-$ whose quality is too low. The other strategy can be taking both $\boldsymbol{y}^+$ and $\boldsymbol{y}^-$ by S2L-RM. This strategy ensures the quality of both responses, but the data may have lower diversity. It is still an open question to answer which strategy is better, especially when talking about the long context RLHF. We leave this as our next phase of study.

4 Conclusion

In conclusion, we introduced Short-to-Long Reward Modeling (`S2L-RM`) as a method to generate preference data for long-context RLHF using short-context reward models. Our approach was validated through assumption testing in short-context scenarios and preliminary RLHF experiments with Best-of-N optimization. The positive results demonstrate the potential of S2L-RM in addressing the challenge of aligning long-context LLMs, paving the way for future work in preference-based tuning for such models.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. Longalign: A recipe for long context alignment of large language models. *arXiv preprint arXiv:2401.18058*, 2024a.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longwriter: Unleashing 10,000+ word generation from long context llms. *arXiv preprint arXiv:2408.07055*, 2024b.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

ChatGLM. Glm long: Scaling pre-trained model contexts to millions. https://medium.com/@ChatGLM/glm-long-scaling-pre-trained-model-contexts-to-millions-caa3c48dea85, 2024. Accessed: 20 September, 2024.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

Zhiyuan Hu, Yuliang Liu, Jinman Zhao, Suyuchen Wang, Yan Wang, Wei Shen, Qing Gu, Anh Tuan Luu, See-Kiong Ng, Zhiwei Jiang, et al. Longrecipe: Recipe for efficient long context generalization in large languge models. *arXiv preprint arXiv:2409.00509*, 2024.

Maurice George Kendall. Rank correlation methods. 1948.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling, 2024.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*, 2023.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Yiqi Wu, Xiaodan Hu, Ziming Fu, Siling Zhou, and Jiangong Li. Gpt-4o: Visual perception performance of multimodal large language models in piglet activity understanding, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

A Appendix

Table 1: Case study. The baseline approach $r(\boldsymbol{q}, \boldsymbol{y})$ has encountered reward hacking in this example, where a response with the highest score is factually inconsistent with the reference response. Our approach, both best and worst, has selected the responses that are factually consistent with the reference response. Although the response from **S2L (Worst)** contains redundant information, it explicitly conveys uncertainty.

Long-form Context: LONGBENCH: A BILINGUAL, MULTITASK BENCHMARK FOR LONG CONTEXT UNDERSTANDING
ABSTRACT Although large language models (LLMs) demonstrate impressive performance for many language tasks, most of them can only handle texts a few thousand tokens long, limiting their applications (... omitted due to space constraints)

Query: Which datasets in LongBench were constructed or labeled from scratch by the authors?

Reference Response:
Five datasets in LongBench were constructed or labeled from scratch by the authors: 1. MultiFieldQA-en: the authors collected documents and articles from different domains and asked three PhD students to label questions and answers for each article. 2. MultiFieldQA-zh: similar to MultiFieldQA-en. 3. PassageCount: The author randomly selects passages from the English Wikipedia, repeats them and disrupts the order, the task requires the model to determine the number of unique passages. 4. PassageRetrieval-en: the author randomly selects passages from the English Wikipedia, and generates a summary of a passage using GPT-3.5-Turbo. The task requires the model to recognize the original passage corresponding to the summary. 5. PassageRetrieval-zh: similar to PassageRetrieval-en, the model uses the contents of the Chinese C4 dataset.

$r(\boldsymbol{q}, \boldsymbol{y})$ **(Best):**
According to the text, 5 datasets in LongBench were created and annotated by the authors, while 16 datasets were either directly extracted from existing datasets or built based on existing datasets and processed to be suitable for long context evaluation. The 5 datasets created and annotated by the authors are not explicitly listed in the text, but they are mentioned as 5̈ are created and annotated by usïn the abstract and 5̈ are created and annotated by the authorsïn the table captions.

S2L (Best):
According to the text, the following 5 datasets in LongBench were constructed or labeled from scratch by the authors: 1. MultiFieldQA-en 2. MultiFieldQA-zh 3. PassageCount 4. PassageRetrieval-en 5. PassageRetrieval-zh

S2L (Worst):
According to the text, the authors constructed or labeled 5 datasets from scratch, which are: * MultiFieldQA datasets in both English and Chinese * PassageCount * PassageRetrieval-en and PassageRetrieval-zh * LCC (Long Code Completion) dataset is sampled from the original Long Code Completion dataset, but it is not clear if the original dataset was constructed or labeled from scratch. However, the authors did sample from it and adapted it for LongBench. * RepoBench-P dataset is adapted from the original RepoBench dataset, but it is not clear if the original dataset was constructed or labeled from scratch.