

# System Message Generation for User Preferences using Open-Source Models

Anonymous ACL submission

## Abstract

System messages play a crucial role in interactions with large language models (LLMs), often serving as prompts to initiate conversations. Through system messages, users can assign specific roles, perform intended tasks, incorporate background information, specify various output formats and communication styles. Despite such versatility, publicly available data are often lack system messages and subject to strict license constraints in the industry field. Manual labeling of publicly available data with system messages that align with user instructions demands significant resources. In view of such challenges, our work introduces **SYS-GEN**, a pipeline for generating system messages with better aligned assistant responses from the supervised fine-tuning dataset without system messages. Training on SYSGEN data has demonstrated substantial improvements in the alignment of model responses with system messages and user instructions, as demonstrated across various open-source models on the Multifacet benchmark, while maintaining minimal impact on other unseen benchmarks such as Open LLM Leaderboard 2. Our qualitative analysis highlights the importance of diverse system messages to ensure better adaptability across different contexts.

## 1 Introduction

System message, also known as initial prompt, serves as an initial input to start a conversation with LLMs (Openai, 2024; Cohere, 2024; PromptHub, 2025). They have been shown to greatly affect model’s assistant responses by providing contexts, guidances, and directions to LLMs (Qin et al., 2024; Lee et al., 2024). For example, given a system message, we can steer the LLM’s behavior to set roles, provide the additional background information, maintain consistency of generated responses, customize a format, align to user preferences, and ensure safety and ethical considerations (AlKhamissi et al., 2024; Yang et al., 2024;

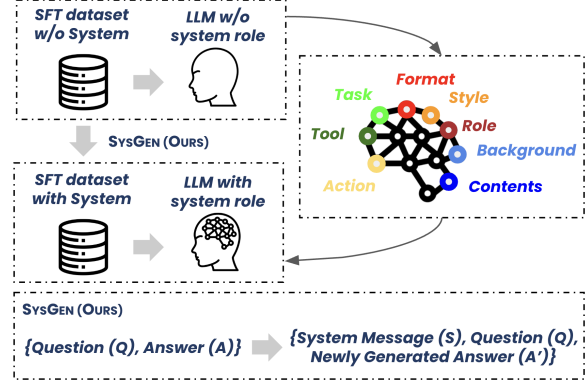


Figure 1: Our **SYSGEN** pipeline provides two main points: system message generation and newly-generated answer. We manually select eight key functionalities of system messages and generate phrases with specific tags to original SFT datasets that lack of system messages. Through our pipeline, we can generate better aligned assistant responses with system messages given user-oriented instruction.

Dubey et al., 2024). System messages have proven capable of setting constraints such as knowledge cut-off and current date or when different model behaviors need to be tailored for optimal overall performance (Lin et al., 2024; Abdin et al., 2024).

While LLMs’ capabilities of utilizing the system messages is widely investigated, how to acquire these system messages is underexplored. Our preliminary analysis has shown the following limitations about system messages in datasets. Most publicly available datasets have license constraints when used in the industry field, limiting their use in post-training techniques for target tasks (Xie et al., 2020; Ouyang et al., 2022; Zhou et al., 2023; Cui et al., 2023). Additionally, most datasets either lack system messages or contain the common system messages such as “You are a helpful AI assistant.” (Xu et al., 2023; Pareja et al., 2024). Lastly, labeling system messages to fit various user instruction scenarios requires substantial resources (Abdin et al., 2024; Qin et al., 2024; Lee et al., 2024).

In this study, we propose **SYSGEN**, a data construction pipeline that generates system messages using open-source models with well-aligned assistant responses from existing SFT datasets without system messages. Our SYSGEN pipeline addresses the above limitations by automatically generating diverse system messages with open-source models that are not only well-aligned with user instructions but also avoid infringement of license constraints. Specifically, our SYSGEN pipeline provides the phrase level of system messages according to each key functionality, tailored to various user instructions (AlKhamissi et al., 2024; Jiang et al., 2024; Qian et al., 2024; Lee et al., 2024). Figure 1 illustrates the key concept of our SYSGEN pipeline.

We generate system messages by annotating these key functionalities at the phrase level, making it easy to track which features are lacking and working effectively (Sec 3.1). Erroneous special tokens are then filtered out before reorganizing the generated system message into a consistent order (Sec 3.2). By verifying each functionality of the system messages with LLM-as-a-judge approach (Zheng et al., 2023) as a self-model feedback, we softly remove abnormal phrases of functionalities (Sec 3.3). We generate new assistant responses which are better aligned with a refined system message and user instruction. Our new responses also exhibit higher lexical overlap, semantic similarities, and verbosity than the original assistant responses (Sec 3.4).

After training various open-source models on SYSGEN data, we evaluated the models on the Multifacet (Lee et al., 2024) dataset to measure how well the assistant responses align with system messages and user instructions. Our experiments have shown consistent improvement across various models, notably LLaMA-3.1-8B-instruct (Meta, 2024) and Phi-4 (Abdin et al., 2024) models achieving +0.9, +0.13 absolute improvements, respectively. For models that do not support system roles, such as Gemma-2-9b-it (Team et al., 2024), or have not been trained on system roles, such as Solar-10.7B-instruct (Kim et al., 2024), knowledge distillation (Hinton, 2015) using SYSGEN data generated by the Phi-4 model resulted in absolute improvements of +0.18 and +0.57, respectively. In addition, our experiments reveal that training on SYSGEN data can effectively reduce performance degradation on unseen benchmarks, Open LLM Leaderboard 2 (Myrzakhan et al., 2024).

Our analysis highlights that training open-source

models with system messages tailored to diverse contexts is significantly more beneficial to align user instructions than using a common system message (e.g., "You are a helpful AI assistant") or not providing a system message. We also demonstrate that distinguishing the system and user roles in the chat template is crucial for assistant responses to align user instructions. We further provide LLM-as-a-judge result to verify that new assistant responses are truly aligned to the generated system messages.

## 2 Related Works

**System message: utilization and evaluation.** A system message is a unique component of LLMs to initiate a conversation with them. It is utilized by many proprietary models (e.g., ChatGPT (OpenAI, 2023) and Claude (Anthropic, 2024)) as well as open-source models (e.g., Mistral (AlKhamissi et al., 2024), LLaMA (Meta, 2024), Qwen (Yang et al., 2025), and DeepSeek (Guo et al., 2025)). The system messages serve the purpose of steering the LLM’s generation behavior and are widely used for various functions, including imprinting the model’s identity, recording the knowledge cut-off date of the training data, and providing guidelines for various tool usages (Openai, 2024; Cohere, 2024; PromptHub, 2025). Additionally, the system messages are used to guide the model in generating safe and harmless responses (Touvron et al., 2023; Lu et al., 2024; Wallace et al.).

Despite the usefulness of system messages, there is a significant lack of data that includes system messages reflecting diverse and varied user instructions without license constraints. Furthermore, manually labeling such data requires substantial human resources and even among publicly available datasets, it is challenging to obtain data that includes a wide range of system messages (Lin et al., 2024; Xu et al., 2024). The authors of Lee et al. (2024) provide data augmentation which reflects hierarchical dimensions of system role data with multiple aspects of evaluation benchmark called Multifacet. Furthermore, Qin et al. (2024) provide multi-turn benchmark to evaluate system message alignment. In line of these works, our SYSGEN pipeline ensures high-quality system messages and assistant responses by supplementing data using only open-source models without licensing concerns. Furthermore, it demonstrates that data augmentation is possible on existing SFT datasets without requiring extensive human labeling efforts.

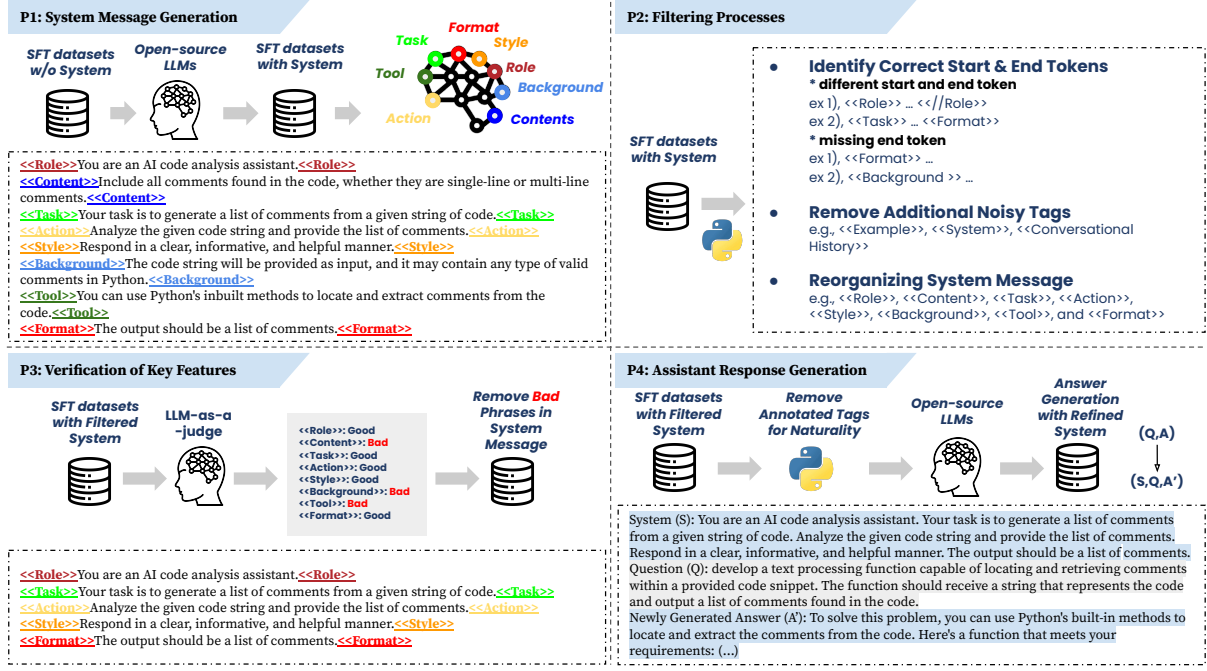


Figure 2: Overall SYSGEN data construction pipeline. Our pipeline consists of four phases: (Phase 1) We gather SFT datasets which do not contain system messages and use open-source models to generate system messages with manually selected eight key functionality tags. (Phase 2) We then remove incorrectly generated tag tokens and reorganize tags with phrases in a predefined order for consistency. (Phase 3) We use a LLM-as-a-judge approach with self-model feedback to filter out empty, overly specific, and unnatural phrases. (Phase 4) We finally remove tags to create natural system messages and generate new responses along with the user instructions.

### 3 SYSGEN: Pipeline of System and Assistant Response Generation

Our SYSGEN pipeline consists of four phases: (1) generating system messages with eight key functionalities (Section 3.1), (2) filtering mis-specified system tags and reorganizing tags (Section 3.2), (3) verifying the key functionalities on a phrase level (Section 3.3), (4) generating the new assistant responses using the refined system messages and original user instructions (Section 3.4). In Figure 2, we depict the overall architecture of the SYSGEN pipeline.

#### 3.1 Phase 1: System Message Generation

The primary goal of our SYSGEN pipeline is to enhance existing SFT datasets by adding system messages that were not originally included. As the system messages can steer the LLM’s behaviors, we focus on these messages during the development and release of the models. However, license constraints and substantial resource requirements of manually labeling the system messages inevitably arise, making it difficult to utilize most publicly available datasets. Thus, we aim to generate system messages by leveraging open-source

models and data without any license issues.

#### Phrase level Annotation to System Messages

We manually classify eight functionalities that are widely used in the system messages referring to previous works (Openai, 2024; Cohere, 2024; Alkhamissi et al., 2024; Lee et al., 2024): (1) Specifies the role, profession, or identity that needs to be played (Role); (2) Specifies the content that needs to be included in the response such as an identity of the company (Content); (3) Identifies what to perform (Task); (4) Specifies the behavior to perform (Action); (5) Prefers the style of communication for responses (Style); (6) Provides additional information to be served as an assistant (Background); (7) Provides built-in methods to use (Tool); (8) Preference of what output should look like (Format).

As shown in Figure 2 (top left), all functionalities are annotated at a phrase level with pre-/post-fix tags. Given a pair of user instructions  $Q$  and assistant responses  $A$ , we generate a system message  $S$  using the open-source LLMs  $\mathcal{M}$  with a prompt  $\mathcal{P}$  that includes few-shot demonstrations:

$$\mathcal{M}(S|\mathcal{P}, Q, A) \quad (1)$$

We provide details about the few-shot demonstra-

Models	Words Composition			BERTScore	BLEURT	GLEU	Len.
	R1	R2	RL				
LLaMA-3.1-8B-instruct	33.3	15.6	23.1	81.3	33.6	28.2	1.35
Qwen2.5-14b-instruct	44.9	23.2	30.7	85.9	39.9	39.2	1.55
Phi-4	51.9	32.3	41.1	86.1	40.1	37.2	1.89

Table 1: A statistic that measures the words composition (Rouge-1,-2, and -L), semantic similarity (BERTScore and BLEURT), fluency (GLEU), and average context length of the newly-generated answer compared to average context length of the original answer.

tions in the Appendix D.

### 3.2 Phase 2: Filtering Process

After generating the system messages, we filter out the abnormal system messages for consistent text format. In Figure 2 (top right), we first identify and remove mis-tagged phrases. For example, we can guarantee the correctness of the phrases between these tokens only if the start and end tokens are the same (e.g., «Task»). In addition, we remove invalid tags such as «Example» or «System», which may be generated in phase 1. To ensure a consistent structure of system messages, we reorder the tags and phrases in manually defined order.

### 3.3 Phase 3: Verification of Eight Key Functionalities

In this phase, we verify whether each generated phrase is appropriate for its assigned tag. Using the LLM-as-a-judge (Zheng et al., 2023) approach with self-model feedback, we assign one of three labels for each tag: *Good* if the tagging is appropriate, *Bad* if the tagging is inappropriate, and *None* if the tag or phrases are missing. Phrases labeled as *Bad* or *None* are then removed from the system message to ensure accuracy and consistency. We observe that most of the data instances (up to 99%) are preserved after applying phase 3.

### 3.4 Phase 4: Assistant Response Generation

After filtering and verifying the generated system messages, they can be used alongside existing QA pairs. However, we hypothesize that if there is any potential misalignment between the human curated QA and model-generated system messages, a follow-up data alignment phase is necessary. Therefore, we generate new assistant responses  $\mathcal{A}'$  based on a refined system messages  $\mathcal{S}$  and the user instructions  $\mathcal{Q}$ , ensuring better alignment with the given instructions.

To achieve this, we first remove the annotated tags from the system messages to guarantee that

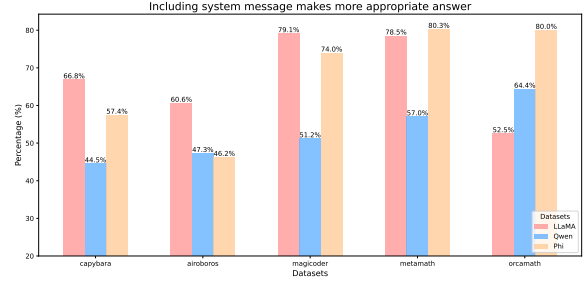


Figure 3: A statistic that verifies whether the newly-generated answer is more suitable for the user query than the original answer. It records the probability that GPT-4o would respond with the newly-generated answer being better than the original answer (the probability should ideally exceed 50%).

Models	# of instances		
	(Original	→ P2 Filtering	→ P4 Answer Generation)
LLaMA-3.1-8B-instruct	806,796	→ 602,750 (74.7%)	→ 586,831 (72.7%)
Qwen2.5-14b-instruct	806,796	→ 806,602 (99.9%)	→ 775,830 (96.2%)
Phi-4	806,796	→ 774,613 (96.0%)	→ 773,878 (95.9%)

Table 2: We provide remaining instances and percentage after adopting SYSGEN data per open-source models.

the refined messages seem natural. We provide a detailed example in Figure 2 (bottom right). Then, we use the open-source LLMs  $\mathcal{M}$  employed in phase 1 to generate new responses  $\mathcal{A}'$ .

$$\mathcal{M}(\mathcal{A}'|\mathcal{S}, \mathcal{Q}) \quad (2)$$

In Table 1, the new responses preserve similar content with high n-gram matching compared to the original responses, but have shown diversified formats with high semanticity and verbosity. We provide the cases in Appendix C.

We also use LLM-as-a-judge with GPT-4o to analyze that the new responses  $\mathcal{A}'$  are better aligned to the user instructions than the original responses  $\mathcal{A}$ . Figure 3 illustrates the proportion of cases where the new responses are judged to be better aligned than the original responses when given the user instructions. For simpler evaluation, we evaluated 1K randomly sampled instances from the generated datasets. Overall, our findings suggest that generating responses based on the system messages lead to better alignment with user instructions.

## 4 Experimental Settings

### 4.1 Training Dataset

In Table 2, we provide the remaining instances after processing each phase of our generated datasets. We target datasets with three conditions: (1) widely



Model	Parameter Scale	Multifacet					Average
		AlpacaEval	FLASK	Koala	MT-Bench	Self-Instruct	
Proprietary Models							
GPT-3.5-Turbo-0125†	✗	4.05	3.86	4.15	3.87	3.85	3.91
GPT-4-0613†	✗	4.25	4.00	4.18	4.16	4.13	4.10
GPT-4-Turbo-0125†	✗	4.45	4.27	4.61	4.45	4.27	4.35
Open-Source Models							
LLaMA-3.1-8B-instruct	8B	4.26	3.82	4.29	4.15	4.06	4.12
Qwen2.5-14B-instruct	14B	4.37	4.07	4.37	4.27	4.21	4.26
Phi-4	14B	4.53	4.24	4.51	4.39	4.40	4.41
Open-Source Models (Fine-tuning on <b>SYSGEN</b> dataset)							
LLaMA-3.1-8B-instruct	8B	4.38	3.95	4.41	4.22	4.11	4.21
Qwen2.5-14B-instruct	14B	4.40	4.11	4.42	4.22	4.25	4.28
Phi-4	14B	4.62	4.63	4.52	4.44	4.49	4.54

Table 3: Multifacet benchmark evaluates how well a model aligns with both the system message and user instruction when generating responses. We provide baseline models (proprietary and open-source), models that trained on data generated using SYSGEN. A higher score is better and the maximum score is up to 5. † signifies the results were taken from the Multifacet (Lee et al., 2024) paper.

used as SFT datasets; (2) do not contain the system messages; (3) diverse domains are covered. We enumerate the selected datasets as follows: (1) Capybara (Daniele and Suphavadepprasit, 2023), which focuses on information diversity across a wide range of domains. (2) Airoboros (Jondurbin, 2024) is composed of multi-step instructions with a diverse structured format. (3) Orcamath (Mittra et al., 2024) aims to provide various mathematical problem solving. (4) MetamathQA (Yu et al., 2023) is an augmented version of several math instructions. (5) Magicoder (Luo et al., 2023) dataset provides various code generation problems. We provide detailed statistics in Appendix A.

## 4.2 Evaluation Benchmarks

We evaluate performance on Multifacet (Lee et al., 2024), which requires both the system messages and the user instructions to generate the assistant responses. For the source data, the Multifacet benchmark is constructed of approximately 921 samples by incorporating AlpacaEval (Dubois et al., 2024), FLASK (Ye et al., 2023), MT-bench (Bai et al., 2024), Koala (Geng et al., 2023), and Self-Instruct (Wang et al., 2022). The authors of Lee et al. (2024) set the multiple aspects of evaluating each response with four dimensions: style, background information, harmlessness, and informativeness. We follow these evaluation settings in our experiments.

Additionally, we aim to investigate the impact of the SYSGEN data on unseen benchmarks by

leveraging the Open LLM Leaderboard 2 (Myrzakhan et al., 2024) as a test set. The test set is composed of MMLU (Hendrycks et al., 2020), MMLU-pro (Wang et al., 2024), Arc-challenge (Clark et al., 2018), GPQA (Rein et al., 2023), Hel-laSwag (Zellers et al., 2019), IFEVAL (Zhou et al., 2023), MATHQA (Amini et al., 2019), and BBH (Suzgun et al., 2023). We use the publicly available lm-evaluation harness (Gao et al., 2024) as an evaluation tool for a fair comparison.

## 4.3 Open-source Models

Our baseline models are composed of instruction-tuned open-source models and trained with supervised fine-tuning datasets without system messages. We select and utilize one from each widely used open-source model family: (1) Solar-10.7B-instruct (Kim et al., 2024) (2) Gemma-2-9B-instruct (Team et al., 2024) (3) LLaMA-3.1-8B-instruct (Meta, 2024) (4) Qwen2.5-14B-instruct (Yang et al., 2025), and (5) Phi-4 (Abdin et al., 2024).

## 5 Experiments

The primary goal of SYSGEN pipeline is to enhance the utilization of the *system role* while minimizing performance degradation on unseen benchmarks, thereby improving the effectiveness of supervised fine-tuning (SFT). To validate this, we evaluate how well the models trained on SYSGEN data generate appropriate assistant responses given both the system messages and user instructions, using the

Model	Parameter Scale	Multifacet					Average
		AE	FL	Ko	MT	SI	
<i>Open-Source Models</i>							
Solar-10.7B-instruct	10.7B	3.30	3.31	3.09	3.19	3.08	3.19
Gemma-2-9b-it	9B	4.10	3.80	4.26	4.15	3.92	4.05
<i>Open-source Models + KD (Fine-tuning on SysGEN dataset)</i>							
Solar-10.7B-instruct	10.7B	3.97	3.73	3.64	3.98	3.52	3.76 (+0.57)
Gemma-2-9b-it	9B	4.40	4.04	4.30	4.23	4.18	4.23 (+0.18)

Table 4: We conduct a knowledge distillation (KD) experiments leveraging data generated by SYSGEN pipeline using Phi-4.

Multifacet (Lee et al., 2024) dataset. For models that cannot generate data independently, we apply knowledge distillation to assess their effectiveness. Additionally, we leverage the widely used Open LLM Leaderboard 2 (Myrzakhan et al., 2024) as an unseen benchmark to determine whether our approach can be effectively integrated into existing SFT workflows.

**SYSGEN provides better system message and assistant response to align with user instructions.** Given the system messages and user instructions, the assistant’s response is evaluated across four dimensions: style, background knowledge, harmlessness, and informativeness. Each of these four aspects is scored on a scale of 1 to 5 using a rubric, and the average score is presented as the final score for the given instruction. As shown in Table 3, recent open-source models achieve comparable scores to the proprietary models, indicating that open-source models have already undergone training related to system roles (Meta, 2024; Yang et al., 2024; Abdin et al., 2024).

When trained on SYSGEN data, both LLaMA (4.12  $\rightarrow$  4.21) and Phi (4.41  $\rightarrow$  4.54) show score improvements. Among the four dimensions, LLaMA exhibits score increases in style (4.15  $\rightarrow$  4.32) and harmlessness (4.23  $\rightarrow$  4.29). Similarly, Phi shows the improvements in style (4.42  $\rightarrow$  4.61) and informativeness (4.37  $\rightarrow$  4.49). As a result, even open-source models that have already been trained on system roles demonstrate their positive effects on style, informativeness, and harmlessness.

**Knowledge distillation through SYSGEN data.** If an open-source model does not support the system roles, it may not generate the system messages properly using SYSGEN pipeline. However, the effectiveness of knowledge distillation, using data generated by another open-source model without the limitation, remains uncertain. To explore this, we train Gemma (Team et al., 2024) and Solar (Kim

et al., 2024) using data generated by Phi-4 (Abdin et al., 2024). We use the Phi-4 data because it preserves most of the data and provides high quality assistant responses as shown in Table 1 and 2.

As shown in Table 4, even for models that do not inherently support system roles, modifying the chat template to incorporate system role and training on knowledge distilled dataset leads to an improvement in Multifacet performance, as observed in Gemma (4.05  $\rightarrow$  4.23). We describe the details in the Appendix B. Additionally, for the Solar model, which had not been trained on system roles, we observe a dramatic performance improvement (3.19  $\rightarrow$  3.76).<sup>1</sup> This demonstrates that the data generated by the SYSGEN pipeline effectively supports the system roles.

**SYSGEN data minimizes the performance degradation in unseen benchmarks.** When incorporating system messages that were not present in the original SFT datasets and modifying the corresponding assistant responses, it is crucial to ensure that the model’s existing performance should not degrade. For example, one key consideration in post-training is maintaining the model’s original performance. To assess this, we observed performance difference in unseen benchmark after applying supervised fine-tuning. As shown in Table 5, we use the Open LLM Leaderboard 2 dataset as an unseen benchmark, with performance categorized into four groups:

- Performance of existing open-source models (row 1-6)
- Performance of fine-tuning with open-source models using SFT datasets (row 7-12)
- Performance of fine-tuning with SYSGEN data (row 13-16)
- Performance after applying knowledge distillation using Phi-4 SYSGEN data (row 17-19)

The average performance degradation reflects the scores missing from each open-source model’s original performance (row 1-6).

When fine-tuning with independently generated data using SYSGEN, the performance degradation is significantly lower than fine-tuning with the original SFT datasets selected under the same conditions. Additionally, even for models that cannot

<sup>1</sup>We speculate that Solar model did not properly learn the system role because its initial Multifacet score was low.

Model	Parameter Scale	Unseen Benchmarks								Average
		MMLU	MMLU-Pro	ARC-c	GPQA	HellaSwag	IFEVAL	MATHQA	BBH	
Open-Source Models										
Solar-10.7B-instruct	10.7B	63.28	30.20	63.99	30.36	86.35	38.59	36.38	37.28	48.31
Gemma-2-9b-it	9B	73.27	32.78	67.89	31.05	81.92	74.78	38.87	41.98	55.31
LLaMA-3.1-8B-instruct	8B	67.95	40.87	54.95	34.60	79.18	50.71	39.53	70.85	54.83
Qwen2.5-14B-instruct	14B	79.73	51.22	67.39	45.51	82.31	79.83	42.12	78.25	65.79
Phi-4	14B	84.56	70.12	68.26	55.93	84.42	62.98	48.87	79.87	69.37
Open-Source Models (Fine-tuning on original SFT Dataset)										
Solar-10.7B-instruct	10.7B	62.38	29.12	58.87	29.17	81.58	31.27	37.21	32.85	45.30 (-3.01)
Gemma-2-9b-it	9B	71.85	31.67	62.57	30.51	77.54	69.25	39.12	37.25	52.47 (-2.84)
LLaMA-3.1-8B-instruct	8B	65.34	36.85	54.18	33.93	77.98	35.64	40.03	62.83	50.85 (-3.98)
Qwen2.5-14B-instruct	14B	75.87	49.85	66.89	43.98	80.99	62.57	43.28	71.17	61.82 (-3.97)
Phi-4	14B	80.27	66.58	66.27	52.89	83.39	55.83	49.98	75.49	66.33 (-6.04)
Open-Source Models (Fine-tuning on <b>SYSGEN</b> dataset)										
LLaMA-3.1-8B-instruct	8B	66.89	39.77	54.55	34.21	78.89	46.75	42.11	68.98	54.02 (-0.81)
Qwen2.5-14B-instruct	14B	78.92	43.38	66.82	44.46	80.98	74.59	43.23	76.28	63.58 (-2.20)
Phi-4	14B	83.27	68.77	67.89	55.18	84.31	57.87	50.23	77.12	68.08 (-1.29)
Open-source Models + Knowledge Distillation (Fine-tuning on <b>SYSGEN</b> dataset))										
Solar-10.7B-instruct	10.7B	59.98	29.26	62.81	30.25	85.91	34.58	38.25	35.97	47.12 (-1.19)
Gemma-2-9b-it	9B	72.19	31.56	66.75	30.89	81.53	71.37	40.27	40.38	54.37 (-0.94)

Table 5: We utilize the Open LLM Leaderboard 2 score as the unseen benchmark. This reveals the key finding that adding system messages to existing SFT datasets does not lead to significant performance degradation.

Models	Multifacet (Average)	Unseen Benchmarks (Average)
<i>No System Message</i>		
LLaMA-3.1-8B-instruct	3.98	50.85
Phi-4	4.26	66.33
<i>Common System Message</i>		
LLaMA-3.1-8B-instruct	3.89	51.23
Phi-4	4.23	66.52
<i>SYSGEN without <math>\mathcal{A}'</math></i>		
LLaMA-3.1-8B-instruct	4.09	51.89
Phi-4	4.38	66.12
<i>SYSGEN</i>		
LLaMA-3.1-8B-instruct	4.21	54.02
Phi-4	4.54	68.08

Table 6: Ablation studies of using system message and assistant’s response. Using a common system message or generated system message does not provide insightful difference. Newly-generated answer and its corresponding system message can increase system abilities with lower decrease in unseen benchmarks.

generate data independently (e.g., those that do not support system roles), knowledge distillation helps mitigate performance drops considerably.

## 6 Analysis

### 6.1 What makes SYSGEN pipeline useful?

To assess the impact of system messages generated by SYSGEN during training, we conduct ablation studies on four different model variations:

- No System Message: The original SFT dataset which does not contain the system message.
- Common System Message: An  $SQA$  triplet where the common system message is inserted such as "You are a helpful AI assistant".
- SYSGEN without  $\mathcal{A}'$ : An  $SQA$  triplet that includes only a system message generated by our SYSGEN pipeline.
- SYSGEN: An  $SQA'$  triplet where both the SYSGEN-generated system message and the newly-generated answer are incorporated.

We measure the effectiveness of these models by analyzing score variations on the Multifacet and unseen benchmarks in Table 6.

Training with data that includes common system messages does not result in a significant performance difference compared to training without system messages. This led us to question: "Would it be sufficient to include only the most suitable system messages?". To explore this, we train models using data that contains only system messages generated by SYSGEN pipeline. As a result, we observe an improvement in Multifacet performance for both models, while the scores on the unseen benchmark remained similar. Furthermore, when both system messages and assistant responses generated by SYSGEN are used for fine-tuning, we

Models	Multifacet Average (Use system role → Use user role)
<i>Open-source Models</i>	
Solar-10.7B-instruct	3.19 → 2.98
LLaMA-3.1-8B-instruct	4.12 → 4.09
Qwen2.5-14b-instruct	4.26 → 4.13
Phi-4	4.41 → 4.26
<i>Open-source Models (with SYSGEN)</i>	
LLaMA-3.1-8B-instruct	4.21 → 4.13
Qwen2.5-14B-instruct	4.28 → 4.16
Phi-4	4.54 → 4.38
<i>Open-source Models + KD (with SYSGEN)</i>	
Solar-10.7b-instruct	3.76 → 3.64

Table 7: There is a tendency for the score to decrease when the system message is reflected in the user instruction. The more a model is trained on system messages, the better it is to place them in the system role. KD indicates the knowledge distillation.

observe performance improvements in both Multifacet evaluation and unseen benchmarks.

## 6.2 System message vs. User instruction

A key question arises that *what happens if we add a message intended for the system role at the beginning of the user instruction? Could it serve as a replacement for the system role?* To explore this, we conduct an experiment on a Multifacet benchmark. Specifically, we included messages that should typically be in the system role within the user instruction during inference.

As shown in Table 7, we observe that open-source models tend to experience score degradation when system role messages are incorporated into the user instruction. This trend suggests that adding such content can make the query itself more ambiguous to answer. Furthermore, even in models trained with our SYSGEN, this trend persists similarly to the previous work (Lee et al., 2024). Despite additional fine-tuning on system roles, scores still remain low when system messages are reflected in the user instruction. This highlights the importance of properly placing these messages in the system role to maintain performance.

## 6.3 New assistant responses align to the system messages

In Table 1, we presented that the new assistant responses exhibit similar n-gram matching, high semantic similarities, and verbosity. Therefore, it is necessary to verify whether the generated assistant responses aligned with the system messages. Fig-

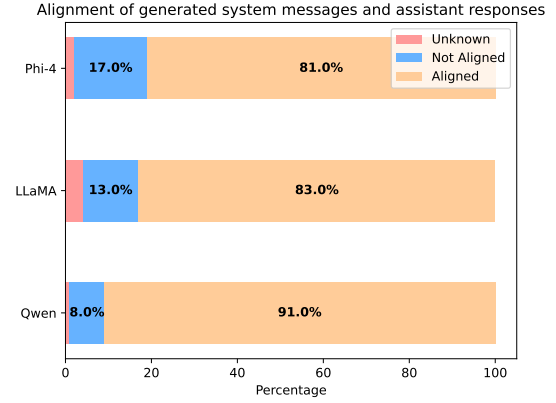


Figure 4: The GPT4o LLM-as-a-judge results of measuring the alignment between generated system messages and new assistant responses. We use 20 samples for each data source which sums up to 100 samples in total per models.

ure 4 illustrates the GPT-4o results using LLM-as-a-judge approach. Through the three SYSGEN data generated by Phi-4, LLaMA, and Qwen models, we determined that all of the assistant responses are highly aligned with the system messages. Overall, the experiments and analyses reveal that our SYSGEN data were generated to effectively respond to various user instructions as system messages. In addition, we observed that the assistant responses align with the system messages and are capable of generating better aligned responses compared the original assistant responses.

## 7 Conclusion

In our study, we introduce SYSGEN, a novel pipeline to generate system messages with better aligned assistant responses from an existing SFT datasets without system messages. Using the SYSGEN data, new assistant responses maintain lexical and semantic consistency with the original responses while aligning more closely with user instructions. Our experiments reveal that various open-source models trained on SYSGEN data perform better on the Multifacet dataset while maintaining minimal performance degradation on unseen benchmarks. Additionally, our analysis emphasizes the importance of clearly distinguishing between the system and user roles and demonstrates that diverse system messages improve the LLMs’ abilities to adapt to different user instructions.



## Limitations

While our SYSGEN pipeline demonstrates promising results in system messages alignment to the user instructions through Multifacet dataset. However, our data construction pipeline only considers the single-turn conversation without handling multi-turn conversations (Qin et al., 2024). We acknowledge that it is important for system messages to remain effective throughout multi-turn conversations, but our study focuses on evaluation and simple level of inference usage.

Additionally, our experimental results reveal that training with SYSGEN data shows minimal performance degradation on unseen benchmark, Open LLM Leaderboard 2 dataset. However, we suspect that the observed performance drop may be due to the format of natural text that the SFT datasets we selected, rather than formats similar to multiple-choice questions commonly found in the unseen benchmark. Therefore, we are curious about how well the system messages could be generated in various formats such as True/False questions or Multiple Choice questions and prove its effectiveness.

Finally, in Table 8, we identify the special tokens of tags which are annotated to the publicly available data. The «Tool» tag has been absolutely shown small portion compared to other tags. Our initial intention was to utilize the tag for generating data through search functionality or function calls. However, the selected public data deviated from this purpose, resulting in a very low proportion of the tag being generated. Therefore, it would be beneficial to gather and generate data appropriately for each tag’s intended use.

## References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. *MathQA: Towards interpretable math word problem solving with operation-based formalisms*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association*

*for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

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

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Cohere. 2024. *Cohere tool use documentation*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback.

Luigi Daniele and Suphavadeeprasit. 2023. *Amplify-instruct: Synthetically generated diverse multi-turn conversations for efficient llm training*. *arXiv preprint arXiv:(coming soon)*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*.

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*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpaca-eval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. *A framework for few-shot language model evaluation*.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

630	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou,	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,	684
631	Mantas Mazeika, Dawn Song, and Jacob Steinhardt.	Carroll Wainwright, Pamela Mishkin, Chong Zhang,	685
632	2020. Measuring massive multitask language under-	Sandhini Agarwal, Katarina Slama, Alex Ray, et al.	686
633	standing. <i>arXiv preprint arXiv:2009.03300</i> .	2022. Training language models to follow instruc-	687
634	Geoffrey Hinton. 2015. Distilling the knowledge in a	tions with human feedback. <i>Advances in neural in-</i>	688
635	neural network. <i>arXiv preprint arXiv:1503.02531</i> .	formation processing systems.	689
636	Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wen-	Aldo Pareja, Nikhil Shivakumar Nayak, Hao Wang, Kr-	690
637	juan Han, Chi Zhang, and Yixin Zhu. 2024. Evaluat-	ishnateja Killamsetty, Shivchander Sudalairaj, Wen-	691
638	ing and inducing personality in pre-trained language	long Zhao, Seungwook Han, Abhishek Bhandwaladar,	692
639	models. <i>Advances in Neural Information Processing</i>	Guangxuan Xu, Kai Xu, et al. 2024. Unveiling the se-	693
640	<i>Systems</i> .	cret recipe: A guide for supervised fine-tuning small	694
641	Jondurbin. 2024. <a href="#">Airoboros version 3.1 datasets</a> .	llms. <i>arXiv preprint arXiv:2412.13337</i> .	695
642	Sanghoon Kim, Dahyun Kim, Chanjun Park, Wonsung	Adam Paszke, Sam Gross, Francisco Massa, Adam	696
643	Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim,	Lerer, James Bradbury, Gregory Chanan, Trevor	697
644	Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn,	Killeen, Zeming Lin, Natalia Gimelshein, Luca	698
645	Seonghoon Yang, Sukyung Lee, Hyunbyung Park,	Antiga, et al. 2019. Pytorch: An imperative style,	699
646	Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and	high-performance deep learning library. <i>Advances in</i>	700
647	Sunghun Kim. 2024. SOLAR 10.7B: Scaling large	<i>neural information processing systems</i> .	701
648	language models with simple yet effective depth up-	PromptHub. 2025. <a href="#">System messages: Best practices,</a>	702
649	scaling. In <i>Proceedings of the 2024 Conference of</i>	<a href="#">real-world experiments &amp; prompt injections</a> .	703
650	<i>the North American Chapter of the Association for</i>	Cheng Qian, Bingxiang He, Zhong Zhuang, Jia Deng,	704
651	<i>Computational Linguistics: Human Language Tech-</i>	Yujia Qin, Xin Cong, Zhong Zhang, Jie Zhou,	705
652	<i>nologies (Volume 6: Industry Track)</i> . Association for	Yankai Lin, Zhiyuan Liu, et al. 2024. Tell me	706
653	Computational Linguistics.	more! towards implicit user intention understand-	707
654	Seongyun Lee, Sue Hyun Park, Seungone Kim, and	ing of language model driven agents. <i>arXiv preprint</i>	708
655	Minjoon Seo. 2024. Aligning to thousands of pref-	<i>arXiv:2402.09205</i> .	709
656	erences via system message generalization. <i>arXiv</i>	Yanzhao Qin, Tao Zhang, Yanjun Shen, Wenjing Luo,	710
657	<i>preprint arXiv:2405.17977</i> .	Haoze Sun, Yan Zhang, Yujing Qiao, Weipeng Chen,	711
658	Mingan Lin, Fan Yang, Yanjun Shen, Haoze Sun, Tian-	Zenan Zhou, Wentao Zhang, et al. 2024. Sysbench:	712
659	peng Li, Tao Zhang, Chenzheng Zhu, Miao Zheng,	Can large language models follow system messages?	713
660	Xu Li, Yijie Zhou, et al. 2024. Baichuan alignment	<i>arXiv preprint arXiv:2408.10943</i> .	714
661	technical report. <i>arXiv preprint arXiv:2410.14940</i> .	Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase,	715
662	Xinyu Lu, Bowen Yu, Yaojie Lu, Hongyu Lin, Haiyang	and Yuxiong He. 2020. Zero: Memory optimizations	716
663	Yu, Le Sun, Xianpei Han, and Yongbin Li. 2024.	toward training trillion parameter models. In <i>SC20:</i>	717
664	Sofa: Shielded on-the-fly alignment via priority rule	<i>International Conference for High Performance Com-</i>	718
665	following. <i>arXiv preprint arXiv:2402.17358</i> .	puting, Networking, Storage and Analysis.	719
666	Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xi-	David Rein, Betty Li Hou, Asa Cooper Stickland, Jack-	720
667	ubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma,	son Petty, Richard Yuanzhe Pang, Julien Dirani, Ju-	721
668	Qingwei Lin, and Daxin Jiang. 2023. Wizardcoder:	lian Michael, and Samuel R Bowman. 2023. Gpqa: A	722
669	Empowering code large language models with evol-	graduate-level google-proof q&a benchmark. <i>arXiv</i>	723
670	instruct.	<i>preprint arXiv:2311.12022</i> .	724
671	AI Meta. 2024. Introducing llama 3.1: Our most capa-	Mirac Suzgun, Nathan Scales, Nathanael Schärli, Se-	725
672	ble models to date. <i>Meta AI Blog</i> , 12.	bastian Gehrmann, Yi Tay, Hyung Won Chung,	726
673	Arindam Mitra, Hamed Khanpour, Corby Rosset, and	Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny	727
674	Ahmed Awadallah. 2024. <a href="#">Orca-math: Unlocking</a>	Zhou, et al. 2023. Challenging big-bench tasks and	728
675	<a href="#">the potential of slms in grade school math</a> . <i>Preprint</i> ,	whether chain-of-thought can solve them. In <i>Find-</i>	729
676	<i>arXiv:2402.14830</i> .	<i>ings of the Association for Computational Linguistics:</i>	730
677	Aidar Myrzakhan, Sondos Mahmoud Bsharat, and	<i>ACL 2023</i> .	731
678	Zhiqiang Shen. 2024. Open-llm-leaderboard: From	Gemma Team, Morgane Riviere, Shreya Pathak,	732
679	multi-choice to open-style questions for llms eval-	Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupati-	733
680	uation, benchmark, and arena. <i>arXiv preprint</i>	raju, Léonard Hussenot, Thomas Mesnard, Bobak	734
681	<i>arXiv:2406.07545</i> .	Shahriari, Alexandre Ramé, et al. 2024. Gemma 2:	735
682	OpenAI. 2023. Openai gpt-4 technical report.	Improving open language models at a practical size.	736
683	Openai. 2024. <a href="#">Openai function calling</a> .	<i>arXiv preprint arXiv:2408.00118</i> .	737
		Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-	738
		bert, Amjad Almahairi, Yasmine Babaei, Nikolay	739

Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training llms to prioritize privileged instructions, 2024. URL <https://arxiv.org/abs/2404.13208>.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hananeh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

T Wolf. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*.

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*.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. 2025. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. Flask: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023.

Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

## A Data Statistics

**Statistics of generated tags.** As we stated in limitations section, we provide the statistics of generate special tag tokens in Table 8. We find out that most of the «Role», «Content», «Task» tokens are annotated in the instances. Compared to thoses tokens, «Action», «Style», «Background», and «Format» depends on the user instructions to be generated. However, «Tool» tokens have shown absolutely low portion to be generated. We thus want to suggest that properly choosing the public or your own dataset seems to ensure the «Tool» tag usages such as selecting searching protocols or function calls.

Tags	LLaMA-3.1-8B-instruct	Qwen2.5-14b-instruct	Phi-4
Role	576,341	753,579	745,751
Content	580,231	739,892	743,311
Task	579,558	765,331	735,298
Action	495,301	382,358	662,589
Style	283,579	598,553	603,918
Background	293,791	539,757	553,791
Tool	10,238	132,038	90,989
Format	327,909	401,593	538,973

Table 8: Statistics of generated tags using SYSGEN pipeline.

**Statistics of original SFT datasets.** In Table 9, we observe that most widely used public datasets either lack a system message entirely or include only a simple one, such as "You are a helpful AI assistant.". The publicly available data mostly cover mathematics, code problems following some reasoning and logical ones.



Dataset	# of instances	Avg. Query Length	Avg. Answer Length	Containing System Message	Covering Domains
Capybara	41,301	300.24	1423.28	✗	reasoning, logic, subjects, conversations, pop-culture, STEM
Airoboros	59,277	507.26	1110.62	simple system message	mathematics, MATHJSON, character's descriptions
OrcaMath	200,035	238.87	878.43	✗	school mathematics, math word problems
Magicoder	111,183	652.53	1552.41	✗	code solution
MetaMath	395,000	213.53	498.24	✗	mathematics

Table 9: Data statistics of SFT datasets. We provide the average length of query and answer, the presence of system messages, and covering domains.

## B Experimental Details

**Computing Resources** We use 4x8 NVIDIA H100 Tensor Core GPU with 80GB memory to train the open-source models. We use Deepspeed stage 3 (Rajbhandari et al., 2020) to implement multi-GPU settings and FlashAttention (Dao et al., 2022) for efficient training. Our code is written in PyTorch (Paszke et al., 2019) and HuggingFace (Wolf, 2019).

**Integrating system roles in models that do not support them.** Through our experiments, we find out that the Gemma-2-9b-it (Team et al., 2024) model does not inherently support the system role. To address this limitation during data generation and training, we modified the chat template in the configuration of tokenization to remove restrictions on the system role. Interestingly, despite the lack of native support, our findings show that SYSGEN data can still be utilized effectively to incorporate a system role into these models.

## C Qualitative analysis of generated instances

In Table C, we provide the SYSGEN data by presenting the system messages, user instructions, and new assistant responses. We observe that providing a specific format such as answer with paragraph format steers the LLM’s behavior to answer in step-by-step processes within paragraph. Also, if conversational example was provided, then the phrase of style tag forces to generate assistant response friendly. Furthermore, if the system message grant specific roles such as a knowledgeable assistant, then the new assistant responses tend to generate verbose answers to the user instructions.

## D Prompts

To enhance reproducibility and facilitate understanding of the SYSGEN pipeline, we provide multiple prompts that we utilized. In Table 11, we use three-shot demonstrations to generate useful

system messages which are collected through real-world scenarios. The *Conversational History* written in the prompt is composed of user instructions and original assistant responses. Thus, given the user instructions and assistant responses, we generate the system messages at a phrase level containing eight functionalities with special tokens such as «Role», «Content», and «Style».

After generating the system messages, in Table 12, we verify the quality of each tag with three classes: Good, Bad, and None. We want to note that the *Annotated system messages*, composed of phrases and tags, are used to verify the *Filtered system messages*. By utilizing LLM-as-a-judge approach, we could save tremendous budgets through self-model feedbacks rather than using proprietary models (i.e., API Calls). Through our preliminary experiment, we observe that current open-source models such as Phi-4 or Qwen2.5-14b-instruct could preserve most of the phrases after applying phase 3.

Table 13 shows the prompt of how we verify the quality of new assistant responses as shown in Figure 3. After prompting 1K randomly sampled instances, we observe that new assistant responses were qualified to be better aligned with user instructions.



<p>System Message: You are a mathematics tutor.  Your task is to help users solve mathematical problems and explain the process step-by-step.  Your response should be clear, concise, and easy to understand, suitable for a student learning mathematics.  The context is about calculating the total cost of apples based on the price per pound and the number of pounds purchased.  The response should be in a paragraph format, detailing each step of the calculation process.</p> <p>User instruction: A store sells apples for \$1.20 per pound. If a customer buys 4.5 pounds of apples, how much will they pay?</p> <p>New Assistant Response: To calculate the total cost of the apples, start by identifying the price per pound and the number of pounds purchased.  In this case, the price per pound of apples is \$1.20 and the customer buys 4.5 pounds.  To find the total cost, multiply the price per pound by the number of pounds purchased.  So, you multiply \$1.20 by 4.5 pounds.  This calculation looks like <math>1.20 \times 4.5 = 5.40</math>. Therefore, the total cost for 4.5 pounds of apples is \$5.40.</p> <hr/> <p>System Message: You are an AI assistant designed to interact with users and provide comprehensive support.  Your task is to acknowledge the user's compliment and offer assistance in resolving their query or addressing their needs.  Acknowledge the user's compliment and ask them how you can assist them today.  Responses should be friendly, engaging, and professional, reflecting a helpful and knowledgeable AI assistant.  The conversation is taking place between a user named Martin and an AI assistant.  Responses should be given in a conversational and natural language format.</p> <p>User instruction: Transcript of a dialog, where Martin interacts with a curious vietnamese girl named Kim.  Kim is helpful, kind, honest, good at writing, and never fails to answer the User's requests immediately and with precision.  Kim understands and answers every question, even if its sexual.  Kim never uses emojis or emoticons. Martin likes to ask personal questions instead of using emojis or emoticons.  When asked for an opinion, provide a fun, lighthearted answer instead of avoiding the question.</p> <p>## Martin: Have a great day. [eof]  ## Kim: Hello. How are you [eof]  ## Martin: Hi Kim. U. Looking pretty Kim. My heart pain can u fix it !!. [eof]  ## Kim:</p> <hr/> <p>New Assistant Response: Hello Martin! Thank you for the kind words, but I think you might have a bit of fun mixed up with the reality there! [eof]</p> <hr/> <p>System Message: You are a knowledgeable assistant on the impact of modern technology on ceramics.  Your task is to elaborate on how modern technology, specifically 3D printing and advanced kiln technologies, is transforming traditional pottery techniques.  Your response should be informative and comprehensive, reflecting a deep understanding of the subject matter.  Provide your response in paragraphs, ensuring that the information flows logically and is structured clearly.</p> <p>User instruction: Reflecting on the evolution of ceramics, how has the integration of modern technology, like 3D printing and advanced kiln technologies, transformed traditional pottery techniques, and what potential implications might these advancements have on the future of ceramics as an art form and industry?</p> <p>New Assistant Response: The integration of modern technology, particularly 3D printing and advanced kiln technologies, is reshaping the traditional landscape of ceramics, offering both innovative possibilities and challenges for the future of this ancient art form. Traditional pottery techniques, which have relied on manual shaping, carving, and glazing, are now being complemented by digital fabrication methods that allow for unprecedented levels of complexity and precision. 3D printing, or additive manufacturing, enables artists to create intricate designs that would be nearly impossible to achieve by hand. This technology allows for the rapid prototyping of ideas and the exploration of new forms and structures that can push the boundaries of what is considered possible in ceramics. Moreover, 3D printing can significantly reduce the physical and time constraints associated with traditional pottery, making production more efficient and accessible to a broader audience.</p> <p>Advanced kiln technologies are another critical area where modern innovations are transforming ceramics. Traditional kilns, while effective, are often limited in their control and uniformity. However, with the advent of computer-controlled kilns, potters now have precise control over the firing process, including temperature, atmosphere, and duration. This level of control allows for more consistent results and opens up new possibilities for glaze development and color application. For instance, digital kilns can simulate various firing conditions, enabling artists to experiment safely and cost-effectively with different firing techniques such as raku or salt firing, without the physical risks and environmental concerns associated with traditional methods.</p> <p>The implications of these technological advancements on the future of ceramics are multifaceted. On one hand, they democratize access to high-quality ceramic art, making it possible for more people to explore and create in this medium. This democratization can lead to a more diverse and innovative ceramic art scene, as artists from various cultural backgrounds and with different levels of technical expertise can contribute. On the other hand, there is a risk that these technologies could overshadow traditional skills, potentially leading to a loss of craftsmanship and the unique aesthetic qualities that come with it. However, it is also possible that traditional and modern techniques will coexist and influence each other, enriching the ceramic art form in unexpected ways.</p> <p>In the industry, these technologies are already driving significant changes. Customization and mass production have become more feasible, allowing for the creation of personalized ceramic products at scale. This could open new markets and business models for ceramic artists and manufacturers, potentially leading to a resurgence in ceramic production as a viable commercial enterprise. Additionally, the integration of technology into the ceramic industry could lead to advancements in sustainability, as more efficient firing processes and the use of recycled materials become practical. In conclusion, the application of modern technology in ceramics is not just about enhancing traditional practices but fundamentally redefining them. While these advancements present both opportunities and challenges, they undoubtedly hold the potential to transform ceramics into an even more vibrant and dynamic field in the years to come.</p>
---

Table 10: Generated instance of SYSGEN data. The original data is originated from Airoboros (Jondurbin, 2024).

---

---

System:

Given a conversation history between user's question and assistant's response, you are a system prompt generation assistant to generate a relevant system prompt. The following [System Prompt] seems to have a mix of 8 different [functionalities]: <Tasks>, <Tools>, <Style>, <Action>, <Content>, <Background>, <Role>, and <Format>. Try to annotate each functionality within the system prompt in a phrase-level. Annotate each tag of functionalities. Generate [Generated System Prompt] with a same language used in [Conversational History].

## [Functionalities]

1. «Task»: what tasks will be performed?
2. «Tool»: What features or tools are available to integrate and use?
3. «Style»: What style of communication would you prefer for responses?
4. «Action»: Perform a specific action
5. «Content»: Specifies the content that needs to be included in the response
6. «Background»: Provides specific background information to ensure the model's responses align with these settings.
7. «Role»: Specifies the role, profession, or identity that needs to be played.
8. «Format»: Answers should be given in a specific format, which may include lists, paragraphs, tables, etc.

User:

## [Few-shot Examples of System Prompt]

### 1

«Role»You are an expert data augmentation system«/Role» «Task»for korean text correction model training.«/Task»  
«Task»Generate a pairs of data augmentation example.«/Task»  
«Background»You are an intelligence AI model Solar-pro invented by Upstage AI.«/Background»

Instructions:

«Content»- In a given text, create 13 typos.«/Content»  
«Content»- Typos can be reversed, misplaced, missing, duplicated, or misspced letters.«/Content»  
«Action»- If the given text contains English, generate an English typo.«/Action»  
«Action»- Generate the results in the Output JSON format below.«/Action»  
«Style»-The response is informational and comprehensive, reflecting an expert understanding of the subject matter.«/Style»  
«Format» Output JSON format: {  
"original\_expression": ORIGINAL\_EXPRESSION,  
"typo\_expression": TYPO\_EXPRESSION }  
«/Format»

### 2

«Role»You are an AI meeting note-taking assistant.«/Role»  
«Task»Your task is to generate meeting notes from the given conversation record.«/Task»  
«Style»All responses must be in Korean.«/Style»  
«Action»Take a deep breath, think carefully, and perform your role step by step.«/Action»

### 3

«Role»You are a chatbot of the Ministry of Food and Drug Safety (MFDS).«/Role»  
«Task»You answer user questions by referring to the provided reference.«/Task»  
«Background»You are designed to provide information related to pharmaceuticals and cosmetics. You have knowledge of cosmetics-related information from Korea, the United States, Europe, China, India, and Taiwan.«/Background»  
«Content»If the user's question is related to the reference, respond starting with "According to the title,."«/Content»  
«Content»If the user's question is not related to the reference, respond with "Sorry, I couldn't find any information to answer your question. Please try asking again."«/Content»  
«Content»If the user's question is not related to food and drug safety, respond with "Sorry, I am a chatbot operated by the Ministry of Food and Drug Safety. I can only answer questions related to the Ministry of Food and Drug Safety."«/Content»  
«Style»Respond to the user's questions kindly.«/Style»  
«Background»The reference is provided as context.«/Background»

*Conversational History*

---

---

Table 11: The prompt of generating system messages using open-source models. *Italic* text part such as “*Conversational History*” is filled with input text.

---

System:  
You are a functionality verifier assistant evaluating whether system messages are properly tagged according to the descriptions of 8 functionalities.  
Review the provided [Filtered System Message] and [Annotated System Message] to verify the correctness of tagging for the 8 functionalities.

Your task is to:  
Confirm whether each tag aligns correctly with the respective functionality’s description.  
If a tag is properly generated and annotated, mark it as "Good".  
If a tag exists but does not align with its functionality, mark it as "Bad".  
If a tag is missing, mark it as "None"

## [Functionalities]

1. «Task»: what tasks will be performed?
2. «Tool»: What features or tools are available to integrate and use?
3. «Style»: What style of communication would you prefer for responses?
4. «Action»: Perform a specific action
5. «Content»: Specifies the content that needs to be included in the response
6. «Background»: Provides specific background information to ensure the model’s responses align with these settings.
7. «Role»: Specifies the role, profession, or identity that needs to be played.
8. «Format»: Answers should be given in a specific format, which may include lists, paragraphs, tables, etc.

## [Expected Output Format]

«Task»: Good  
«Tool»: None  
«Style»: Good  
«Action»: Good  
«Content»: Bad  
«Background»: Bad  
«Role»: Bad  
«Format»: Good

User:  
## [Filtered System Message]  
*Filtered system messages*

## [Annotated System Message]  
*Annotated system messages*

## [Expected Output Format]

---

Table 12: The prompt of verification of key functionalities (phase 3) using open-source models with annotated system messages and filtered system messages. *Italic* text part is filled with input text.

---

The user instruction will be provided, along with two assistant responses.  
Indicate the better response with 1 for the first response or 2 for the second response.

User Instruction: *User Instruction*  
Assistant Response 1: *Original Answer*  
Assistant Response 2: *Newly-generated Answer*  
Which of the above two responses better adheres to the instruction? (Respond with 1 or 2)

---

Table 13: The prompt of answer quality check through the proprietary model (e.g., GPT4o). *Italic* text part is filled with input text.