M2Lingual: Enhancing Multilingual, Multi-Turn Instruction Alignment in Large Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Instruction finetuning (IFT) is critical for aligning Large Language Models (LLMs) to follow instructions. Numerous effective IFT datasets have been proposed in the recent past, but most focus on rich resourced languages such as English. In this work, we propose a diverse, task taxonomy guided, fully synthetic **M**ultilingual, **M**ulti-turn evoled instruction finetuning dataset, called **M2Lingual**, to better align LLMs on a diverse set of languages and tasks. **M2Lingual** contains a total of 182K IFT pairs that are built upon diverse seeds collected from Aya collection and Aya dataset covering 70 languages, 19 NLP tasks and general instruction-response pairs. LLMs finetuned with **M2Lingual** substantially outperform the majority of existing multilingual IFT datasets. Importantly, LLMs trained with **M2Lingual** consistently achieves competitive results across wide variety of evaluation benchmarks compared to existing multilingual IFT datasets that enable LLMs performance in only one or a few subset of the benchmarks. Specifically, LLMs finetuned with **M2Lingual** achieve strong performance on multi-turn evaluation benchmarks such as MT-Bench and across wide-variety of multilingual tasks such as XQuAD, MGSM, TyDiQA, MLQA, XNLI and XLSUM. We show efficacy of **M2Lingual** across LLMs with different sizes, especially smaller LLMs with 1.8B size which benefit massively from our dataset. Lastly, we present key analyses to highlight importance of each synthesis step of **M2Lingual**.[1]

1 Introduction

Large language models (LLMs) have achieved remarkable success [1, 20, 21, 45, 53, 43], largely fueled by the availability of a wide variety of high-quality instruction fine-tuning (IFT) datasets [49, 48, 42, 31, 52, 4, 58]. However, most IFT data curation efforts focus on English or widely spoken languages, leaving low-resource languages and multilingual datasets underrepresented [52]. Prior multilingual datasets can be either categorized into machine translated, human generated and human-AI generated datasets. Datasets like MultiAlpaca [5], Bactrian-X [29] and PolyLM [50] utilize machine translations and self-instruct [48] to generate instruction-response (IR) pairs in multiple languages. However, naïve machine translation of English instructions may not capture native or regional knowledge alignment from different languages [47]. Human-generated datasets like Aya [40] and Open Assistant [24] preserve regional knowledge alignment and cultural contexts, making them higher quality compared to translated datasets. However, gathering multilingual annotations in different languages from native speakers (on a large scale) is expensive, time consuming, and prone to annotator errors [40]. Finally, human-AI generated datasets like LMSYS-1M [57], ShareGPT, Vicuna [6], and WildChat [55] involve a human interacting with an AI assistant to gather data. Although such datasets are relatively cheaper to gather compared to human-generated ones, they still

[1] We will release **M2Lingual** data, pipeline and the finetuned models after decision.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

come with several challenges. These include privacy issues, moderate complexity instructions, and necessary legal regulatory constraints[17].

As summarized in Table 1, most of the above datasets like Aya, Bactrian-X, MultiAlpaca are single-turn only, which limits a model's ability to engage in long, multilingual conversations. Additionally, IFT datasets with multi-million scale like the Aya collection [40] containing $513M$ IR pairs or XP3 [33] having $75+M$ pairs can be expensive to finetune LLMs or analyze their different qualitative aspects.Additionally, many multilingual datasets do not include diverse NLP tasks and general instructions across low resource languages, often containing very simple instructions, thus limiting their effectiveness in training strong multilingual instruction following LLMs.

To address these shortcomings, we present **M2Lingual**, a diverse *Multilingual*, *Multi-turn* IFT dataset which is fully synthetic, containing machine generated $182K$ instructions, and covering 70 languages. The dataset is built upon seed samples from a) the human generated Aya dataset, where general IR pairs are annotated by regional, native language speakers, and b) seeds from Aya collection that contain IR pairs from 17 diverse NLP tasks. Unlike previous IFT datasets that use self-instruct mechanism and machine translation to generate data in more languages, **M2Lingual** is constructed with a task-specific taxonomy guided evolve (denoted as *Evol*) conditions [51] to generate new IR pairs from the seed samples in each language. The *Evol* taxonomy covers a diverse range of NLP tasks, regional dialects and slang, resulting in instructions that are diverse, detailed, more complex, and longer in length. Furthermore, to improve LLMs in engaging multilingual conversations we define a multi-turn *Evol* taxonomy for generating conversational IR pairs. The multi-turn (MT) taxonomy covers a wide variety of possible subsequent user instructions as discussed in Section 3, shown in Figure 2. The proposed data enrichment taxonomy for curating new complex & diverse instruction and multi-turn conversations is generic, and can be extended to any monolingual or multilingual data. We also ensure balanced *Evol* generations across languages, creating IR pairs equally for all 70 languages.

In evaluations, we conduct experiments across several multilingual NLP benchmarks and a multi-turn benchmark called MT-Bench [51], translated in 8 languages. We empirically demonstrate the effectiveness of **M2Lingual** by comparing with finetuning several LLMs from different family and sizes with existing multilingual IFT datasets. Our results show that **M2Lingual** leads to best or second best performance across a) several multilingual evaluation benchmarks datasets, and b) multi-turn MT-Bench evaluations. On the other hand, existing IFT datasets show competitive results but only in a few subset of evaluations while performing poorly in other evaluations.

The key contributions from our work are as follows:

1. We present **M2Lingual**, a fully synthetic multilingual, multi-turn IFT dataset of $182K$ IR pairs that lead to best performance in both multilingual evaluation benchmarks and complex multi-turn evaluation dataset MT-Bench. **M2Lingual** is synthesized using a data enrichment taxonomy focused on adding instruction-specific and multi-turn specific evolve [51] complexities. Our data enrichment taxonomies and steps used in **M2Lingual** synthesis can be easily extended to other languages, monolingual settings, and any instruction seeds.
2. **M2Lingual** contains (roughly) equal distribution of IFT pairs across 70 languages ensuring strong performance improvements in multiple languages but notably in low resource languages. Additionally, smaller LLMs like QWEN-1.8B show massive improvements when finetuned with **M2Lingual**, highlighting its usefulness with more accessible models.
3. We present several key ablation studies to highlight the impact of every data enrichment and synthesis steps used in **M2Lingual** generation. We show that adding instruction-task specific complexities improves average performance across several multilingual evaluation benchmarks whereas adding multi-turn specific evols leads to strong improvements on MT-Bench.

2 Related Work

Pretraining LLMs is computationally expensive, and due to abundance vs scarcity of corpus in different languages [27, 34, 26, 23], the majority of pretraining is done in high resource languages like English. This often leads to LLMs performing much better in high resource languages [32, 34] compared to low resource languages. Multilingual instruction finetuning has proven to be a relatively cost effective solutions for improving multilingual performance of LLMs, especially for low resource languages [37, 5, 46]. While several IFT datasets have been introduced in the recent past, less focus has been given on synthesizing multilingual IFT datasets limiting progress in various languages.

Dataset	Size	Multi turn?	Langs	Resource Level		Task specific?	General instructions?	Translated dataset?	Fully synthetic?
				Low	High				
OpenAssistant	10K convs	✓	35	3	32	✗	✗	✗	✗
Aya Dataset	200K IR pairs	✗	70	37 (1)	32	✗	✓	✗	✗
MultiAlpaca	52K IR pairs	✗	12	0	12	✓	✓	✓	✓
Bactrian-X	3.4M IR pairs	✗	52	15(1)	36	✗	✓	✓	✓
ShareGPT	94K convs	✓	45	4 (2)	39	✓	✗	✗	✗
WildChat	1.04M convs	✓	74	21 (3)	50	✗	✗	✗	✗
M2Lingual	182K convs	✓	70	37 (1)	32	✓	✓	✗	✓

Table 1: Comparison of multilingual IFT datasets with **M2Lingual**. Resource level classification taken from NLLB [13]. Languages not found in the NLLB table are counted as low, in parentheses.

Instruction finetuning datasets are often created from pool of numerous NLP tasks (e.g., flanT5, supernatural instructions) [8, 38, 49], machine generated (e.g., self-instruct [48]), human expert annotated (e.g., LIMA [58]) and crowd-sourced or cached from real users chat (e.g., LMSYS, WildChat) [55]. A few synthetic IFT datasets [29, 42] have leveraged LLMs such as GPT-4 for generating IR pairs via self-instruct as a relatively affordable alternative. Although very effective, a few works have highlighted issues in data generation process of self-instruct [4, 16] where generations can be considered a bit uncontrolled. For example, Alpaca [41] uses self-instruct to generate $52K$ instructions from 175 seeds, where overlapping and noisy IFT pairs have been reported [4, 58]. In contrast, techniques like WizardLM [51] that generate new instructions by adding complexity (or *Evol*) on input seed instructions, have generations which are controlled by the set of *Evol* conditions, ensuring diverse generations. For example, as shown in fig. 1, the *Concretize Evol* in left block creates a new IFT pair with a more complex but concrete python question. Furthermore, generating templated datasets for specific NLP tasks (e.g., XP3 [33]) may not contain complex and diverse inputs for aligning instruction following of LLMs. Thus, inspired from WizardLM, **M2Lingual** contains *Evol* IFT pairs generated from a set of diverse IFT seeds. Additionally, we also create a multi-turn *Evol* taxonomy (fig. 2) which is used to generate multi-turn IR pairs resulting in a diverse, complex conversational IFT set within **M2Lingual**. Features of some of the existing IFT datasets are summarized in table 1.

3 Methodology

In this section, we detail the three main synthesizing steps of **M2Lingual**. Step 1 (Section 3.1) involves selection of diverse multilingual seed data. In our work, we select seed samples from two different sources of Aya — Aya dataset and Aya Collection, both of which receive high average approval ratio by human annotators [40]. Step 2 (Section 3.2) and 3 (Section 3.3) correspond to our novel *Evol* taxonomy based data enrichment techniques. Specifically, in Step 2 we create an NLP task specific *Evol* taxonomy [51] and generate new IR pairs using these *Evol* conditions. In Step 3, we first create an *Evol* taxonomy for multi-turn or conversational IR pairs, and then use these *Evol* for generating new conversational, multi-turn IR pairs. Figure 1 captures an overview of each of these generation steps used in synthesis of **M2Lingual**.

3.1 Seed Selection

Our first seed source, Aya dataset, contains general IR pairs written by native speakers/annotators which enables capturing region-specific language nuances and cultural contexts. We randomly select 100 prompts for each of the 70 languages resulting in 7000 seed samples from the Aya dataset. Our second seed source, Aya collection, covers 19 different NLP tasks where each task has parallel examples in 113 different languages. To ensure a proper balance of the number of examples across all languages, we only focus on 70 languages of the Aya dataset. We exclude two NLP tasks from Aya collection - 1) text simplification as it requires rewriting a complex or a simplified version of a sentence which is already supported by our evols, and 2) multilingual event entity task as Aya collection does not have a consistent format for this task. Finally, for each task in the Aya collection, we randomly sample 6 examples per language, resulting in $6 * 70 * 17 = 7140$ seed examples. We select 6 random samples per task per language to ensure balanced amount of seed samples from Aya collection when compared to the seeds from Aya dataset. Thus, our final seed contains $7000 + 7140 = 14140$ samples.

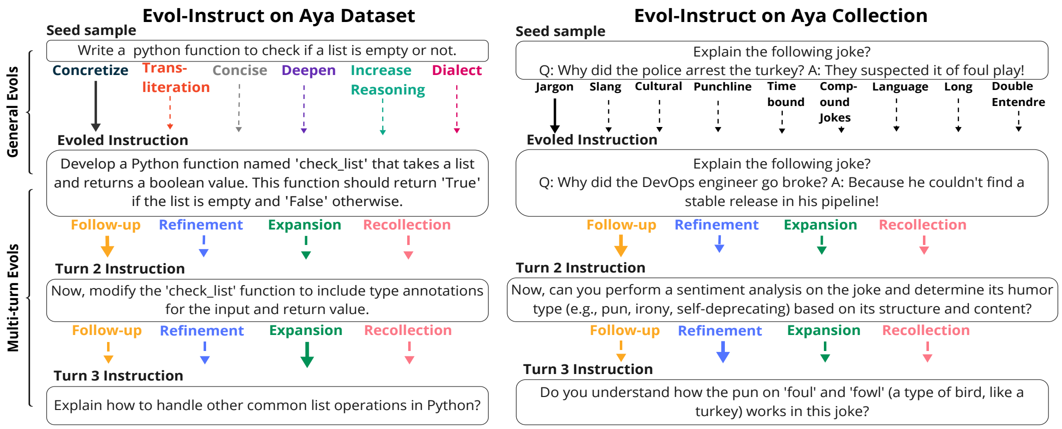

Figure 1: Walk-through examples for data synthesis of **M2Lingual**. In Step 1, seeds are selected from Aya dataset (left figure) and Aya collection (right figure). In Step 2, the task specific *Evol* taxonomy is used for generating new, complex and evolved instructions. General *Evol* are used for seeds from Aya dataset and NLP task specific (see Joke explanation task evols) *Evol* are used for Aya collection seed. Finally, in Step 3, multi-turn instructions are generated on top of the new evolved instruction generated from Step 2.

3.2 Task Guided *Evol*

Though the IR pairs in the seed data from Aya cover a wide variety of NLP tasks, overall they are direct and less intricate. To enhance the instruction following abilities of LLMs, especially for complex tasks and prompts, we employ *Evol*-Instruct [51] on our selected seed instructions as the second data synthesis step. *Evol*-Instruct generates a more complex instructions using *Evol* conditions over the provided seed instruction [51]. The generic *Evol* conditions used in the original work[2] are not always applicable to a wide variety of downstream NLP tasks such as multi-hop question answering, joke explanation, etc. Furthermore, generic *Evol* conditions also provide weak or broader guidance for new IFT pair generations, especially for seed datasets like ours which cover diverse NLP tasks.

Dataset	Seed	Evoled	Multi-turn
Aya Dataset	7000	37803	36969
Aya Collection	7140	57145	34426
Total	**14140**	**94948**	**71395**
Avg Instruction	49.60	107.71	356.81
Avg Response	56.79	62.16	87.60

Table 2: **M2Lingual** IR pairs. Avg Instruction and Response show avg no of tokens.

To address this, we create a taxonomy of *Evol* conditions covering general instructions (for seed from Aya dataset) and each of the NLP tasks (for seeds from Aya collection) as shown in Figure 2. Specifically, we create 6 *Evol* conditions enhancing multi-lingual features focused solely on general instructions. We then leverage GPT-4 to come up with 9 different *Evol* for each NLP task. These NLP task specific *Evol* ensure that we create *Evol* conditions separately for a particular task. Figure 2 shows all the NLP task names and the corresponding evols.

The *Evol* prompts to GPT-4 for each task in Figure 2 are presented in detail in Appendix 9.3. We apply these task-specific *Evol* on our selected seeds:

- The seed samples from the Aya dataset are generic instructions therefore we apply the 6 *generic evols* (Figure 2) to each seed sample. The 6 different evols ensure that the new instruction is more complex, challenging and captures all the multilingual variations, nuances and complexities of different languages. This results in $7K \times 6 = 42K$ samples from the seeds in Aya dataset.
- The seed samples from Aya collection have 17 NLP tasks which are shown in the top block of Figure 2 along with their corresponding 9 evols. For each seed from a particular task, we apply its corresponding 9 evols resulting in a total of $7140 \times 9 = 64260$ instructions from Aya collection.
- Upon manual inspection of the generated instructions across both Aya dataset & collection, we observe that some of the instructions generated using GPT-4 have repetitive long sequences and n-grams. Therefore, following [18, 14], we filter instructions with frequent n-grams. Some requests to GPT-4 also return a time out. The final datasets thus contains $95K$ instructions, $37K$ from the Aya dataset and $57K$ from Aya collection as shown in Table 2.

[2] `https://github.com/lcw99/evolve-instruct/blob/main/evolve.py`

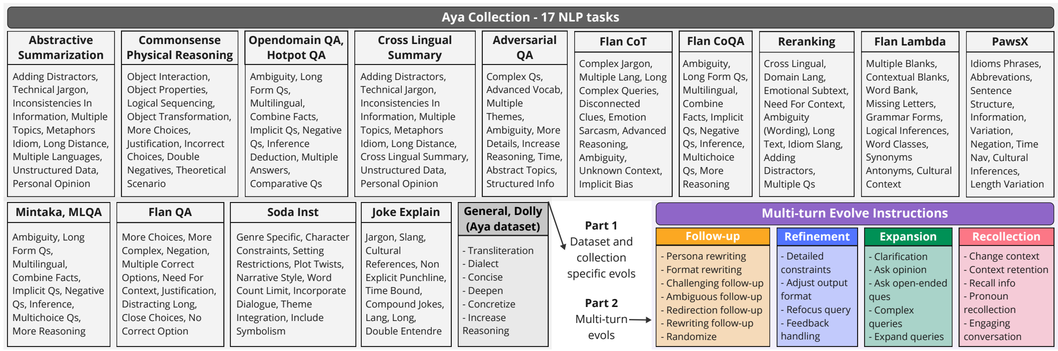

Figure 2: Taxonomy of *Evol* applied towards creating **M2Lingual**. Part 1 includes evols on the general Aya dataset as well as the task-specific Aya collection data, after which Part 2 multi-turn *Evol*are applied for creating multiple turns in the conversation.

3.3 Generating Multiple Turns

As the final step 3 in synthesis of **M2Lingual**, we generate multiple user-assistant turns from the task-evolved instructions produced in the previous step. A conversation between a user and an AI assistant broadly can be categorized into four categories [25] — *Follow-up*, *Refinement*, *Expansion*, and *Recollection*. However, these categories are generic and do not encompass the full complexity and fine-grained variety of conversational interactions. To address this, we introduce a multi-turn taxonomy comprising 21 distinct variations of dialogue that expand upon these four categories. These 21 detailed taxonomy with the 4 categories improve coverage of possible variations in continuing a conversation, thus ensuring an engaging interaction between a user and an assistant [25]. We also ensure that subsequent instructions are generated in the language of the initial instruction by explicitly prompting GPT-4. The last block in Figure 2 shows all the 21 variations. The multi-turn *Evol* prompts to GPT-4 are shown in Appendix 9.4. We convert the instructions to multi-turn conversations with following steps:

1. We use the prompt specified in Appendix 9.4 and replace the *{instruction}* with the task-evoled generated instructions from the previous step (i.e., Step 2), *{follow_up_type}* with one of the 21 dialogue variations in 9.4, and *{language}* with the *Evol* instruction language. We then pass it to GPT-4 n times to generate the next user instruction.
2. For all the generated instructions, we generate subsequent response turns using GPT-4 using the entire conversation history. To mitigate the potential impact of topic drift from the prolonged conversations [54], we restrict the number of subsequent instructions or multi-turns to $<= 4$.
3. We generate conversations for all the evoled instructions from the Aya dataset, resulting in 36K conversations. For Aya collection, we pick a balanced subset of size 35K across all tasks and languages and generate conversations. After applying the same post-processing steps as mentioned in Section 3.2, we end up with 70K conversations.

In total, **M2Lingual** contains 182K IR pairs, with the exact sizes from different steps of **M2Lingual** synthesis shown in Table 2.

4 Experiments

We conduct experiments across *three* model families & *five* model sizes — Mistral-7B [20], LLaMA-3-8B [45] and QWEN-4B [3]. Furthermore, to demonstrate the effectiveness of our dataset across different model scales, we fine-tune both a larger model, LLaMA-2-13B [44], and a smaller model, QWEN-1.8B [3]. To evaluate how well the datasets work with instruction-tuned models, we also experiment with Mistral-Instruct-7B.

4.1 Baseline Datasets

We use *six* different multilingual datasets as baselines for comparison: 1) the top ranked conversation trees from **Open Assistant** [24], 2) **Aya** [40], 3) self-instruct dataset **MultiAlpaca** [50], 4) machine translated **Bactrian-X** [29] derived from Alpaca-52k [41] and Dolly-15k [12], 5) the **ShareGPT** [3] collection, and 6) **WildChat** [55].

[3] https://sharegpt.com/

For a fair comparison with WildChat, we use 200K non-English conversations, ensuring the same language proportions, and downsampled 60K English conversations, resulting in a total of 260K conversations. Similarly for Bactrian-X, we sample 1M IR pairs ensuring the same language proportion as of the original dataset.

Additional Baselines To highlight the importance of each step in our data curation process, we consider several ablations as baselines. Specifically we conduct experiments by training models using 1) only **Seed** samples, 2) seed samples with the generated evols (**Seed + Evol**) and 3) seeds, evols and the generated multi-turn conversations (**Seed + Evol + MT**). Finally, to see whether adding parallel data (PD) helps in improving the over model's performance, we collect 60K from the Aya collection and train a baseline by augmenting the PD with our full dataset (**Seed + Evol + MT + PD**).

4.2 Training

All training is performed on 8 A-100 80GB NViDIA GPUs [7], with the Axolotl[4] framework. We used Mistral tags [20] for finetuning all models. We use a batch size of 64, max seq length 8192, learning rate of 5×10^{-6}, Adam optimizer [22] with a cosine scheduler and 10 warmup steps. We reserve a 5% validation split, and train all the models until validation loss convergence. We compute the loss only on the targets using fp16 training.

4.3 Evaluation

Multilingual benchmarks. We utilize the EleutherAI evaluation framework [15] for consistent comparisons. We evaluate the performance of different multilingual datasets on the following tasks:

- *Question Answering (QA)*: We focus on 3 multilingual QA datasets 1) XQUAD [2] with QA across 11 languages, 2) TyDiQA [9] which has human generated QA in 11 languages and 3) MLQA [28] with QA in 7 languages. While QA data requires short answer phrases, conversational IR pairs might lead to longer answer span generation. Hence, we use 3 in-context examples to get the right output format for LLMs. In the interest of time, we keep the number of examples per language to 100 for XQUAD and MLQA, and 1000 for TyDiQA. We use the validation set for XQUAD and test set for TyDiQA & MLQA, and compute the standard F1-score.
- *Summarization*: We use the XLSUM [19] dataset and focus on 6 languages - Arabic, English, Spanish, French, Japanese and Russian. We restrict the total number of examples to 100 and prompt the model to generate a summary in the same language as the context. We look at the ROUGE_L [30] & BLEU [35] scores for comparison.
- *Classification*: We focus on XNLI [11] and XCOPA [36] with 15 and 11 languages respectively in a zero-shot setting. We compute the accuracy (Acc) by looking at the log-likelihood assigned to the ground truth answer on the validation set.
- *Multilingual math word problems*: We use MGSM [39], a grade-school math benchmark that translates GSM8K [10] to 10 different languages. Similar to QA tasks, we use 3 in-context examples and compute the exact match (EM) with the ground truth answer.

Translated MT-Bench. To evaluate the conversation and instruction following ability of multilingual models across a wide array of tasks and languages, we translate MT-Bench [56]. MT-Bench comprises of 80 multi-turn questions across 8 domains. The models are required to respond to an initial and a follow-up question and GPT-4 assesses the model's responses on a scale of 1 to 10 (10 being the best), with the overall score being the mean over the two turns. We translate it into 9 different languages with professional linguists to ensure high quality evaluation. We modify the judge prompt to include the language of the question asked at each turn, and additionally instruct GPT-4 to make sure the responses are in the same language as the question asked. We report the average scores across all 80 examples for each language and also report the average MT-Bench score across all languages.

Low-resource Languages. We evaluate models on 6 low-resource languages, including Hindi, Urdu, Thai, Tamil, Bengali and Gujarati using the same aforementioned procedure. Since finding native annotators for these low-resource languages might be difficult, we leverage GPT-4 for translations.

5 Results

Consistent best results: As observed in table 3 and 4, **M2Lingual** leads to best or amongst the top results in both *multi-turn evaluations* (Table 3) and *several NLP task comprised multilingual evaluation benchmarks* (Table 4). On the other hand, the majority of baseline datasets show competi-

[4] https://github.com/OpenAccess-AI-Collective/axolotl

Model	Dataset	MT-EN	MT-FR	MT-IT	MT-JP	MT-ES	MT-DE	MT-NL	MT-PT	MT-AVG
Mistral-7B	Open Assistant	6.72	5.87 (5.90)	6.04	4.19	5.87	5.82	4.97	6.01	5.66
	MultiAlpaca	5.45	4.90 (5.22)	4.63	3.76	5.01	4.66	4.51	4.65	4.77
	Bactrian-X	5.60	5.35 (5.26)	5.46	4.82	5.24	5.53	4.96	5.31	5.25
	ShareGPT	7.04	5.93 (5.70)	5.42	4.75	5.83	6.00	5.27	5.92	5.80
	WildChat	7.02	6.46 (6.77)	6.68	5.50	6.71	**6.43**	**6.51**	**6.89**	6.53
	Aya	6.43	5.42 (5.39)	4.97	3.37	5.45	5.37	4.94	5.12	5.18
	Seed	6.01	5.15 (5.14)	5.35	3.44	5.07	5.98	4.62	4.91	5.04
	Seed + Evol	6.33	5.44 (5.30)	5.46	4.74	5.88	5.61	5.40	5.78	5.56
	Seed + Evol + MT (**M2Lingual**)	**7.13**	**6.75 (6.81)**	**6.9**	**5.70**	**6.81**	6.39	6.34	6.46	**6.54**
	Seed + Evol + MT + PD	5.85	5.75 (5.39)	5.60	4.86	5.81	5.73	5.32	5.74	5.55
LLaMA-3-8B	Open Assistant	6.26	5.15 (5.03)	4.95	4.08	5.26	4.87	5.01	5.48	5.12
	MultiAlpaca	4.96	4.60 (5.09)	4.22	3.30	4.76	4.18	4.32	4.27	4.41
	Bactrian-X	6.27	5.73 (5.77)	5.73	4.83	5.95	5.34	5.41	5.90	5.66
	ShareGPT	7.07	6.17 (5.76)	6.43	5.40	6.10	6.07	5.82	6.13	6.10
	WildChat	**7.20**	**6.74 (6.96)**	6.78	**6.35**	6.86	6.60	6.58	6.72	**6.75**
	Aya	5.95	5.01 (4.50)	5.41	3.86	5.27	4.93	4.66	4.95	4.95
	Seed	4.38	3.55 (3.75)	3.56	2.68	3.52	3.42	3.45	3.54	3.54
	Seed + Evol	6.95	6.41 (6.50)	6.22	5.41	6.35	6.11	5.90	5.27	6.12
	Seed + Evol + MT (**M2Lingual**)	7.17	6.55 (6.82)	**6.86**	6.26	**6.95**	**6.65**	6.58	**6.81**	**6.74**
	Seed + Evol + MT + PD	7.00	6.60 (6.75)	6.61	6.27	6.58	6.55	**6.93**	6.59	5.99

Table 3: MT-Bench evaluations in different languages for LLaMA-3-8B-base and Mistral-7B-base. Performance of Canadian French variant is reported in brackets for MT-FR. Best scores are denoted in bold with dark green while 2nd best scores are highlighted with light green. *Seed* denotes 15.1K seed IR pairs; *Seed + Evol* denotes addition of *Evol* IR pairs over seeds totaling $15.1 + 94.9 = 110.5K$ IR pairs. *Seed + Evol + MT* denotes addition of multi-turn data resulting in $110.5K + 71.5K = 182K$ IR pairs which is **M2Lingual**. *Seed + Evol + MT + PD* denotes adding of 60K machine translated parallel data taken from Aya collection.

tive results only on select evaluation settings. Specifically, on *multi-turn evaluations*, conversational IFT datasets like ShareGPT and WildChat lead to competitive performances but all of the other baseline datasets have low MT-Bench scores. This suggests importance of including multi-turn IR sets within IFT datasets. **M2Lingual** achieves top MT-Bench score in 5 languages and 2nd best in remaining 3 languages, outperforming all of the baseline datasets overall. Similarly, as shown in Table 4, **M2Lingual** also leads top performance across 4 out of 7 *multilingual NLP task evaluation benchmarks* and 2nd best results with very close performance to the best score in remaining 3 benchmarks. It is worth noting that evaluations on classification tasks such as XCOPA and XNLI show very minimal performance variations across all IFT datasets which has been shown in other works as well [29]. In generation tasks (MGSM, XLSUM, including QA tasks MLQA, XQuAD, TyDiQA), **M2Lingual** leads to better results over all the baseline datasets. For instance, with Mistral-7B base model, our proposed **M2Lingual** outperforms the second best baseline (Bactrian-X) by 2.13 and 1.98 F1 points on MLQA and XQuAD respectively.

Model	Dataset	XQUAD	TyDiQA	MLQA	XLSUM		MGSM	XNLI	XCOPA
		F1	F1	F1	ROUGE$_L$	BLEU	EM	Acc	Acc
Mistral-7B	Open Assistant	67.99	54.22	53.64	10.86	0.85	16.05	42.74	56.73
	MultiAlpaca	67.99	64.44	55.69	10.9	1.59	10.41	42.18	58.91
	Bactrian-X	71.91	66.63	60.27	3.30	0.20	17.14	**43.91**	58.64
	ShareGPT	66.33	56.97	50.78	3.31	0.288	11.32	41.13	56.09
	WildChat	72.55	64.27	59.53	3.91	0.41	**18.41**	43.11	58.00
	Aya	70.46	66.95	57.47	**12.5**	**2.01**	13.86	41.78	59.00
	Seed	72.52	65.89	59.33	11.53	1.72	16.95	43.28	57.64
	Seed + Evol	71.01	65.04	57.47	9.8	1.37	14.23	43.00	57.55
	Seed + Evol + MT (**M2Lingual**)	**74.53**	**67.57**	**62.40**	10.42	1.38	15.38	42.12	**59.55**
	Seed + Evol + MT + PD	68.79	62.62	60.00	9.92	1.37	16.45	42.36	59.00
LLaMA-3-8B	Open Assistant	64.38	52.65	47.08	9.38	1.21	17.36	46.17	**63.82**
	MultiAlpaca	75.08	64.49	59.01	**10.98**	**1.45**	10.68	**46.93**	63.55
	Bactrian-X	69.57	56.45	58.51	8.39	1.28	22.86	46.90	62.18
	ShareGPT	56.98	58.48	43.43	3.53	0.40	25.32	45.93	63.00
	WildChat	63.15	59.88	63.16	5.52	0.76	26.36	46.88	62.27
	Aya	75.14	59.60	53.14	10.38	1.39	22.09	45.64	63.55
	Seed	**77.27**	68.57	60.01	9.92	**1.45**	17.18	46.02	62.82
	Seed + Evol	76.17	**69.89**	63.09	8.96	1.23	28.00	46.38	61.36
	Seed + Evol + MT (**M2Lingual**)	75.91	67.84	**63.50**	8.87	1.25	27.36	46.18	62.55
	Seed + Evol + MT + PD	76.69	59.24	60.02	9.84	1.37	**29.00**	46.37	62.09

Table 4: Evaluations of LLaMA-3-8B-base & Mistral-7B-base in different tasks. Same notations as in Table 3

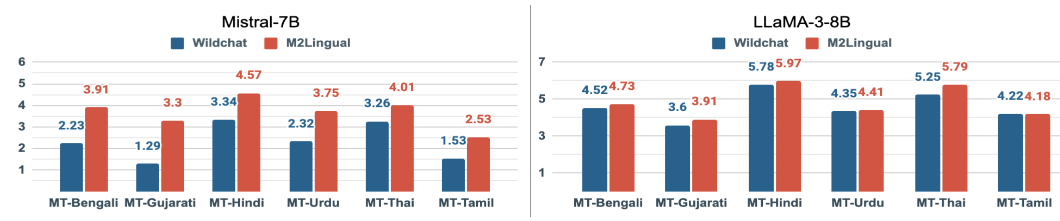

Figure 3: Results on Low Resource translated MT-Bench for Mistral-7B and LLaMA-3-8B

Impact on different LLMs: We evaluate Mistral-Instruct-7B to highlight the impact of multilingual IFT datasets on pre-instruction finetuned models. **M2Lingual** leads Mistral-Instruct-7B to achieve best performance in 5 of 8 MT-Bench language evaluations and 5 of the 7 multilingual evaluation benchmarks as shown in Tables 5 and 8 respectively. Interestingly, the improvements from **M2Lingual** in Mistral-Instruct-7B over baseline datasets is consistently higher when compared to Mistral-7B-base (Table 4) in all of the multilingual QA tasks, MGSM, and XCOPA. We also evaluate QWEN-4B model to showcase results from smaller LLM from different model family. We observe similar findings as QWEN-4B finetuned with **M2Lingual** achieves competitive results in both MT-Bench and multilingual evaluation datasets. Another interesting observation is that improvements seem relatively higher for QWEN-4B model using **M2Lingual** when compared to Mistral-7B and LLaMA-3-8B models, highlighting the usefulness of our proposed data on moderate sized LLMs.

Ablation of M2Lingual: As observed in our empirical studies in Tables 4, 5, 6, 7, generating *Evol* data and appending it with seeds helps improve performance consistently in all evaluation benchmarks, highlighting impact of *Evol*. For instance, on LLaMA-3-8B, our proposed *Evol* improves performance significantly by 10.82 points on MGSM task as compared to the Seed data. Adding multi-turn IFT pairs specifically helps boost performance in MT-Bench evaluations substantially across all of the languages with the most significant gain of 1.31 points on French for Mistral-7B model. Adding multi-turn data also helps consistently in multilingual benchmark evaluations as shown in Tables 4 and 5. *Evol + MT* and *Evol* provides 1.50 and 0.98 points of MT-AVG performance gain over *Seed-only* IFT data on Mistral-7B model. These findings reinforce the benefits of adding multi-turn *Evol* IFT pairs.

Model	Dataset	XQUAD F1	TyDiQA F1	MLQA F1	XLSUM ROUGE$_L$	XLSUM BLEU	MGSM EM	XNLI Acc	XCOPA Acc	MT-Avg
QWEN-4B	Open Assistant	53.63	45.30	46.34	4.15	0.29	17.50	38.52	58.45	3.47
	MultiAlpaca	51.81	53.51	40.26	8.9	1.0	12.1	38.3	58.40	2.93
	Bactrian-X	46.70	42.79	42.2	7.1	0.8	18.6	38.3	57.70	3.80
	ShareGPT	41.86	28.20	36.03	4.58	0.43	16.95	37.83	**58.55**	3.80
	WildChat	53.18	49.18	42.81	5.23	0.56	19.27	**38.74**	58.18	**4.29**
	Aya	54.00	52.14	48.28	**10.91**	**1.31**	16.50	37.59	57.73	3.43
	Seed	**66.55**	**58.09**	48.25	10.65	0.65	15.36	37.59	58.00	2.47
	Seed + Evol	52.24⋆	52.50	49.87	8.50	1.12	20.77	38.36	57.91	3.79
	Seed + Evol + MT (**M2Lingual**)	49.12⋆	47.53	**50.36**	8.30	1.02	**21.36**	38.37	58.36	4.23
	Seed + Evol + MT + PD	57.76	51.97	43.24	9.64	1.21	19.36	37.95	57.91	4.24
Mistral-Instruct-7B	Open Assistant	61.33	59.28	53.27	9.62	1.43	19.00	43.91	58.09	5.58
	MultiAlpaca	63.76	63.05	51.09	11.51	1.80	13.18	**44.70**	58.18	4.74
	Bactrian-X	70.5	64.8	50.60	9.14	1.35	17.91	42.23	57.25	5.98
	ShareGPT	44.53	49.5	40.45	3.31	0.38	17.36	42.13	56.73	6.11
	WildChat	61.53	53.1	52.60	6.31	0.56	21.00	41.86	57.75	6.62
	Aya	69.9	66.43	57.27	**12.58**	**2.05**	16.36	42.84	58.60	5.20
	Seed	68.78	61.54	56.11	12.45	2.04	18.27	43.23	58.45	3.92
	Seed + Evol	**72.87**	68.43	55.43	12.51	1.33	22.00	42.51	58.09	6.48
	Seed + Evol + MT (**M2Lingual**)	71.41	69.44	58.33	9.57	1.51	19.82	42.37	**59.45**	**6.64**
	Seed + Evol + MT + PD	70.04	**69.67**	**59.13**	9.06	1.46	**22.27**	42.92	57.82	6.56

Table 5: Evaluations of QWEN-4B & Mistral-Instruct-7B in different tasks and MT-Bench score averaged across languages. Please see table 8 in appendix for MT-Bench score in each language. ⋆ in XQUAD, TyDiQA scores for QWEN-4B show exception cases where outputs had repeated noisy patterns in multiple runs resulting in low scores.

Low-resource languages: We also compare **M2Lingual** with our most competitive baseline dataset WildChat on MT-Bench in low-resource languages. As shown in Figure 3, **M2Lingual** consistently leads to much higher MT-Bench scores in majority of the low-resource languages highlighting that existing multi-turn IFT datasets created from cached user chats may have poor coverage of low-resourced languages. On the other hand, our synthetically generated **M2Lingual** has uniform coverage of all the 70 languages in terms of number of IFT sets.

6 Additional Analysis

Effect of IFT datasets on different sized LLMs. In addition to 4B, 7B, and 8B sized LLMs shown in Tables 4, 5, 6, 7, we also study impact of our IFT datasets on a smaller LLM (QWEN-1.8B) and a larger LLM (LLaMA-2-13B). As shown in Table 6, on QWEN-1.8B LLM, **M2Lingual** leads to even higher performance for various tasks when compared to our strongest baseline WildChat with the most significant improvements of 13.64 and 22.24 points on MGSM and TyDiQA tasks respectively. Similar findings for the LLaMA-2-13B model highlights the effectiveness of our proposed **M2Lingual** across various sized LLMs.

Model Name	Dataset	MT-EN	MT-FR	MT-IT	MT-JP	MT-ES	MT-DE	MT-NL	MT-PT	MT-Avg	MGSM	MLQA	TyDiQA
Qwen-1.8B	WildChat	4.99	2.75 (2.74)	2.08	1.72	2.88	1.92	1.54	2.68	2.59	8.00	29.39	42.42
	MultiAlpaca	3.97	1.93 (1.99)	1.74	1.44	1.87	1.86	1.52	1.91	2.03	7.45	19.30	33.38
	M2Lingual	**6.20**	**4.55 (4.25)**	**3.85**	**3.27**	**4.40**	**4.11**	**3.32**	**4.51**	**4.27**	**21.64**	**38.24**	**64.66**
LLaMA-2-13B	WildChat	6.64	6.25 (5.89)	5.98	5.10	6.20	6.10	5.82	5.99	6.00	9.95	53.69	60.14
	MultiAlpaca	5.09	4.35 (4.55)	4.35	3.52	4.47	4.54	4.69	4.62	4.46	7.80	48.74	59.46
	M2Lingual	**6.47**	**6.40 (6.20)**	**6.13**	**5.35**	**6.18**	**5.94**	**5.87**	**6.17**	**6.08**	**11.95**	**54.64**	**64.66**

Table 6: Evaluations of QWEN-1.8B and LLaMa-2-13B for highlighting impact on different sized LLMs.

Importance of *Evol*. We selected $15.1K$ seeds from Aya dataset and Aya collection as discussed in section 3.1. We generated *Evol* and multi-turn IR pairs from these seed but as an alternative, more data can also be sampled from Aya. To highlight benefits from *Evol* and generating multi-turn IR pairs, we sample $94.9K$ more IR pairs from Aya collection and Aya dataset, making the total seed size as **M2Lingual** of $110.5K$. As shown in Table 7, simply sampling more seed IFT pairs from Aya achieves low performance, whereas having the same number *Evol*IR pairs from **M2Lingual** leads to much higher performance in MT-Bench and MGSM (2.05 and 6.68 points respectively). It is worth noting that MLQA being reading comprehension QA data requires short answer phrases for exact match. IR pairs within **M2Lingual** are longer (Table 2) and conversational which often lead to longer answer span generation for reading comprehension task. Hence, we utilized 3-shot setting to get the right output format from LLMs in our QA evaluation experiments.

Model	Data	MT-Avg	XQUAD	TyDiQA	MLQA	XLSUM	MGSM	XNLI	XCOPA
Mistral-Instruct-7B	Aya-seeds(110.5K)	4.59	71.40	68.00	**57.69**	**14.08/2.45**	15.32	40.77	57.55
	Seed + Evol (110.5K)	**6.64**	**72.87**	**68.43**	55.43	12.51/1.33	**22.00**	**42.51**	**58.09**

Table 7: Performance comparison of **M2Lingual** vs Aya-seeds data of same size.

7 Conclusion

We build **M2Lingual**, a multilingual, multi-turn IFT dataset that leads to top performances across several multilingual evaluation benchmarks. Our work presents two IFT data enrichment techniques, namely 1) taxonomy based instruction-task specific *Evol*, and 2) multi-turn *Evol* for generating a diverse, conversational multilingual IFT dataset. **M2Lingual** contains roughly same number of IR pairs for 70 languages, resulting in substantial performance improvements in low resource languages. **M2Lingual**-also strongly improves multilingual performance of different sized LLMs ranging from 4B, 7B, 8B, and 13B parameters but in particular leads to massive improvements of small LLMs with 1.8B parameter size. Thus, **M2Lingual** presents a strong societal impact specifically for underrepresented or low-resourced languages. Additionally, the massive performance gains with **M2Lingual** on smaller 1.8B parameter LLMs also contributes towards improving accessibility to the wider community.

8 Limitations and Ethical Considerations

As future work and limitations of our work, **M2Lingual** can be extended to more than 3 turns to generate longer conversational IFT data, although this would be computationally expensive. Similarly, **M2Lingual** can be extended to more number of NLP tasks and languages in future work. In this work, we select seeds from Aya which does not contain specific flags for toxic, harmful, or offensive speech [40], but report low risk. We conducted manual inspection of a few generated IFT pairs from *Evol*, and did not find any harmful IFT data, but future work includes filtering **M2Lingual** by automatic safety tools.

References

[1] OpenAI Josh Achiam et al. "GPT-4 Technical Report". In: 2023. URL: `https://api.semanticscholar.org/CorpusID:257532815`.

[2] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. "On the cross-lingual transferability of monolingual representations". In: *arXiv preprint arXiv:1910.11856* (2019).

[3] Jinze Bai et al. "Qwen Technical Report". In: *ArXiv* abs/2309.16609 (2023). URL: `https://api.semanticscholar.org/CorpusID:263134555`.

[4] Lichang Chen et al. "Alpagasus: Training a better alpaca with fewer data". In: *arXiv preprint arXiv:2307.08701* (2023).

[5] Pinzhen Chen et al. "Monolingual or Multilingual Instruction Tuning: Which Makes a Better Alpaca". In: *The 18th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 2024, pp. 1–10.

[6] Wei-Lin Chiang et al. "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality". In: *See https://vicuna. lmsys. org (accessed 14 April 2023)* (2023).

[7] Jack Choquette et al. "NVIDIA A100 Tensor Core GPU: Performance and Innovation". In: *IEEE Micro* 41.02 (2021), pp. 29–35.

[8] Hyung Won Chung et al. "Scaling instruction-finetuned language models". In: *Journal of Machine Learning Research* 25.70 (2024), pp. 1–53.

[9] Jonathan H Clark et al. "Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 454–470.

[10] Karl Cobbe et al. "Training verifiers to solve math word problems". In: *arXiv preprint arXiv:2110.14168* (2021).

[11] Alexis Conneau et al. "XNLI: Evaluating cross-lingual sentence representations". In: *arXiv preprint arXiv:1809.05053* (2018).

[12] Mike Conover et al. *Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM*. 2023. URL: `https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm` (visited on 06/30/2023).

[13] Marta R Costa-jussà et al. "No language left behind: Scaling human-centered machine translation". In: *arXiv preprint arXiv:2207.04672* (2022).

[14] AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. "Octopus: A Multitask Model and Toolkit for Arabic Natural Language Generation". In: *arXiv preprint arXiv:2310.16127* (2023).

[15] Leo Gao et al. *A framework for few-shot language model evaluation*. Version v0.0.1. Sept. 2021. DOI: `10.5281/zenodo.5371628`. URL: `https://doi.org/10.5281/zenodo.5371628`.

[16] Sreyan Ghosh et al. "A Closer Look at the Limitations of Instruction Tuning". In: *arXiv preprint arXiv:2402.05119* (2024).

[17] Waris Gill et al. "Privacy-Aware Semantic Cache for Large Language Models". In: *arXiv preprint arXiv:2403.02694* (2024).

[18] Han Guo et al. "Efficient (soft) q-learning for text generation with limited good data". In: *arXiv preprint arXiv:2106.07704* (2021).

[19] Tahmid Hasan et al. "XL-sum: Large-scale multilingual abstractive summarization for 44 languages". In: *arXiv preprint arXiv:2106.13822* (2021).

[20] Albert Q Jiang et al. "Mistral 7B". In: *arXiv preprint arXiv:2310.06825* (2023).

[21] Albert Q Jiang et al. "Mixtral of experts". In: *arXiv preprint arXiv:2401.04088* (2024).

[22] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[23] Tom Kocmi et al. "Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet". In: *Proceedings of the Eighth Conference on Machine Translation*. 2023, pp. 1–42.

[24] Andreas Köpf et al. "Openassistant conversations-democratizing large language model alignment". In: *Advances in Neural Information Processing Systems* 36 (2024).

[25] Wai-Chung Kwan et al. "MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models". In: *arXiv preprint arXiv:2401.16745* (2024).

[26] Hugo Laurençon et al. "The bigscience roots corpus: A 1.6 tb composite multilingual dataset". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 31809–31826.

[27] Teven Le Scao et al. "Bloom: A 176b-parameter open-access multilingual language model". In: (2023).

[28] Patrick Lewis et al. "MLQA: Evaluating cross-lingual extractive question answering". In: *arXiv preprint arXiv:1910.07475* (2019).

[29] Haonan Li et al. "Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation". In: *arXiv preprint arXiv:2305.15011* (2023).

[30] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: `https://aclanthology.org/W04-1013`.

[31] S. Longpre et al. "The Flan Collection: Designing Data and Methods for Effective Instruction Tuning". In: *Proceedings of the 40th International Conference on Machine Learning*. 2023, pp. 22631–22648.

[32] Niklas Muennighoff et al. "Crosslingual Generalization through Multitask Finetuning". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 15991–16111.

[33] Niklas Muennighoff et al. "Crosslingual generalization through multitask finetuning". In: *arXiv preprint arXiv:2211.01786* (2022).

[34] Thuat Nguyen et al. "Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages". In: *arXiv preprint arXiv:2309.09400* (2023).

[35] Kishore Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: `10.3115/1073083.1073135`. URL: `https://aclanthology.org/P02-1040`.

[36] Edoardo Maria Ponti et al. "XCOPA: A multilingual dataset for causal commonsense reasoning". In: *arXiv preprint arXiv:2005.00333* (2020).

[37] Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. "Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations". In: *arXiv preprint arXiv:2308.14186* (2023).

[38] Victor Sanh et al. "Multitask Prompted Training Enables Zero-Shot Task Generalization". In: *International Conference on Learning Representations*. 2021.

[39] Freda Shi et al. "Language models are multilingual chain-of-thought reasoners". In: *arXiv preprint arXiv:2210.03057* (2022).

[40] Shivalika Singh et al. "Aya dataset: An open-access collection for multilingual instruction tuning". In: *arXiv preprint arXiv:2402.06619* (2024).

[41] Rohan Taori et al. *Stanford Alpaca: An Instruction-following LLaMA model*. `https://github.com/tatsu-lab/stanford_alpaca`. 2023.

[42] Rohan Taori et al. *Stanford alpaca: An instruction-following llama model*. 2023.

[43] Gemini Team et al. "Gemini: a family of highly capable multimodal models". In: *arXiv preprint arXiv:2312.11805* (2023).

[44] Hugo Touvron et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023).

[45] Hugo Touvron et al. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023).

[46] Ahmet Üstün et al. "Aya model: An instruction finetuned open-access multilingual language model". In: *arXiv preprint arXiv:2402.07827* (2024).

[47] Bin Wang et al. "SeaEval for Multilingual Foundation Models: From Cross-Lingual Alignment to Cultural Reasoning". In: *arXiv preprint arXiv:2309.04766* (2023).

[48] Yizhong Wang et al. "Self-Instruct: Aligning Language Models with Self-Generated Instructions". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 13484–13508.

[49] Yizhong Wang et al. "Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022, pp. 5085–5109.

[50] Xiangpeng Wei et al. “Polylm: An open source polyglot large language model”. In: *arXiv preprint arXiv:2307.06018* (2023).

[51] Can Xu et al. “WizardLM: Empowering large pre-trained language models to follow complex instructions”. In: *The Twelfth International Conference on Learning Representations*. 2023.

[52] Shengyu Zhang et al. “Instruction tuning for large language models: A survey”. In: *arXiv preprint arXiv:2308.10792* (2023).

[53] Susan Zhang et al. “Opt: Open pre-trained transformer language models”. In: *arXiv preprint arXiv:2205.01068* (2022).

[54] Zhuosheng Zhang and Hai Zhao. “Advances in multi-turn dialogue comprehension: A survey”. In: *arXiv preprint arXiv:2103.03125* (2021).

[55] Wenting Zhao et al. “WildChat: 1M ChatGPT Interaction Logs in the Wild”. In: *arXiv preprint arXiv:2405.01470* (2024).

[56] Lianmin Zheng et al. “Judging llm-as-a-judge with mt-bench and chatbot arena”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[57] Lianmin Zheng et al. “Lmsys-chat-1m: A large-scale real-world llm conversation dataset”. In: *arXiv preprint arXiv:2309.11998* (2023).

[58] Chunting Zhou et al. “Lima: Less is more for alignment”. In: *Advances in Neural Information Processing Systems* 36 (2024).

9 Appendix

9.1 MTbench score from QWEN-4B and Mistral Instruct-7B

Model	Dataset	MT-EN	MT-FR	MT-IT	MT-JP	MT-ES	MT-DE	MT-NL	MT-PT	MT-Avg
QWEN-4B	Open Assistant	5.95	3.49 (3.66)	2.84	2.38	3.88	2.73	2.46	3.23	3.47
	MultiAlpaca	4.74	3.29 (2.88)	2.65	1.90	3.15	2.56	2.08	2.90	2.93
	Bactrian-X	5.88	3.84 (4.03)	3.25	2.66	3.85	3.49	2.77	3.90	3.80
	ShareGPT	5.89	3.92 (4.02)	3.39	3.13	4.20	2.97	2.55	3.72	3.80
	WildChat	6.27	4.49 (4.81)	3.83	3.20	4.38	3.83	3.11	4.27	4.29
	Aya	5.24	3.45 (3.74)	2.96	2.24	3.77	3.08	2.44	3.51	3.43
	Seed	4.60	2.68 (2.63)	2.09	1.59	2.43	2.18	1.67	2.03	2.47
	Seed + Evol	5.81	3.86 (4.03)	3.00	2.82	4.24	3.35	2.53	3.68	3.79
	Seed + Evol + MT	6.01	4.67 (4.62)	3.55	3.36	4.48	3.83	2.89	4.02	4.23
	Seed + Evol + MT + PD	5.95	4.35 (4.53)	3.92	3.39	4.38	4.10	3.30	4.23	4.24
Mistral-Inst 7B	Open Assistant	6.76	5.74 (6.07)	5.73	3.78	5.84	5.91	4.99	5.60	5.58
	MultiAlpaca	5.90	4.83 (4.82)	4.66	3.25	5.01	4.57	4.84	4.71	4.74
	Bactrian-X	7.06	5.96 (6.02)	6.22	4.53	6.25	6.09	5.81	6.15	5.98
	ShareGPT	6.84	6.34 (6.20)	5.84	4.61	6.51	6.10	6.06	6.25	6.11
	WildChat	7.39	6.77 (6.53)	6.737	5.64	6.503	6.80	6.39	6.95	6.62
	Aya	5.83	5.32 (5.78)	5.45	3.61	5.39	5.06	5.28	5.32	5.20
	Seed	4.85	4.28 (4.24)	3.98	2.44	3.98	3.71	3.85	4.03	3.92
	Seed + Evol	7.20	6.24 (6.56)	6.40	5.55	6.83	6.41	6.51	6.57	6.48
	Seed + Evol + MT	7.47	6.70 (6.50)	6.71	5.75	6.91	6.52	6.37	6.83	6.64
	Seed + Evol + MT + PD	7.15	6.60 (6.75)	6.73	5.68	6.76	6.28	6.66	6.60	6.56

Table 8: MT-Bench evaluations in different languages for QWEN-4B and Mistral-Instruct-7B.

9.2 Examples of Generated Evols and conversations

Step	Content	Evol
Seed Sample	What is the longest-running Broadway show ever?	
evoled (Turn 1) Instr.	Which Broadway show holds the record for the longest run? Please answer briefly.	Concise
GPT-4 Response	The Phantom of the Opera holds the record for the longest run on Broadway.	
Turn 2	Now, break down the following multi-part question and answer each part: Who wrote this show? What are some of its most popular songs? When did it first premiere on Broadway?	Expanding Queries Change Context
GPT-4 Response	1. Who wrote this show? The Phantom of the Opera was composed by Andrew Lloyd Webber. The lyrics were written by Charles Hart and Richard Stilgoe. 2. What are some of its most popular songs? Some of its most popular songs include "The Phantom of the Opera," "Music of the Night," "All I Ask of You," and "Masquerade." 3. When did it first premiere on Broadway? The Phantom of the Opera first premiered on Broadway on January 26, 1988.	

Table 9: Conversation example from **M2Lingual**

9.3 Prompts for Evol-instruct

For Dolly, HotpotQA and MLQA we use evols from generic, OpenQA and Mintaka respectively.

Step	Content	Evol
Seed Sample	Formulate an answer to this elaborate question: Where is the regional airline based that Jim French is the former chairman and CEO of?	
evoled (Turn 1) Instr.	Formulate an answer to this elaborate question: Identify the base location of the regional airline for which Jim French served as the former chairman and CEO, and explain the significance of this location in the context of the airline's regional operations.	Combine Facts
GPT-4 Response	Jim French served as the former chairman and CEO of Flybe, a British regional airline. The base location of Flybe was Exeter International Airport in Devon, England. The significance of Exeter International Airport as the base location for Flybe lies in its strategic positioning within the United Kingdom. Exeter is situated in the southwest region of England, which allowed Flybe to efficiently connect various regional destinations across the UK and Europe. This location provided a central hub for operations, facilitating shorter flight routes and quick turnaround times, which are essential for regional airlines aiming to offer frequent and reliable services. Additionally, having a base in Exeter helped Flybe tap into a market with fewer competing airlines, allowing it to establish a strong regional presence and foster economic growth within the southwest of England by improving connectivity and accessibility.	
Turn 2	Please adjust the output to provide the answer in a concise text-only format.	Adjust Output Format
GPT-4 Response	Jim French served as the former chairman and CEO of Flybe, a British regional airline based at Exeter International Airport in Devon, England. The significance of this location lies in its strategic positioning in the southwest of England, enabling efficient regional connectivity across the UK and Europe and fostering economic growth in the region.	
Turn 3	Now, rephrase your response as if you were an enthusiastic travel blogger sharing fascinating details with your readers.	Persona Rewriting Follow Up
GPT-4 Response	Hey travel enthusiasts! Did you know that Jim French, the former chairman and CEO of Flybe, once led this fantastic British regional airline based at Exeter International Airport in beautiful Devon, England? It's such a cool spot! Nestled in the southwest of England, Exeter's prime location made it perfect for Flybe to connect countless regional destinations across the UK and Europe. This strategic hub wasn't just about convenience; it played a crucial role in boosting regional connectivity and economic growth. So, next time you're flying through Exeter, remember the incredible role it played in the world of regional air travel!	

Table 10: Conversation from **M2Lingual**

Step	Content	Evol
Seed Sample	Given the following statement, predict the tags for each word: Daniel Smith daniel.smith2@durham.ac.uk Durham University, United Kingdom	
evoled (Turn 1) Instr.	Identify the appropriate tags for each element in this data: Daniel Smith daniel.smith2@durham.ac.uk Durham University, United Kingdom. Please provide a brief response.	Concise
GPT-4 Response	Name: Daniel Smith Email: daniel.smith2@durham.ac.uk Institution: Durham University Country: United Kingdom	
Turn 2	Now, please reformat the output to display only the tags and associated elements in a structured JSON format.	Adjust Output Format
GPT-4 Response	`{ "Name": "Daniel Smith", "Email": "daniel.smith2@durham.ac.uk", "Institution": "Durham University", "Country": "United Kingdom" }`	

Table 11: Conversation from **M2Lingual**

Step	Content	Evol
Seed Sample	Finish the following sentence with the best choice: To clean away the excess plastic from the edge of a project, you can Choices: - Use a belt sander to remove the spare plastic - Use a coffee grinder to remove the spare plastic Answer:	
evoled (Turn 1) Instr.	Complete the sentence with the most appropriate choice: To prevent the documents from flying around in a windy environment, you can Choices: - Use a heavy paperweight to keep them in place - Use a feather to keep them in place Answer:	Object Properties
GPT-4 Response	Use a heavy paperweight to keep them in place	
Turn 2	Now, generate a paragraph explaining the pros and cons of using a heavy paperweight to keep documents in place in windy conditions.	Challenging Follow Up
GPT-4 Response	Using a heavy paperweight to secure documents in a windy environment has several advantages. First, the weight of the paperweight can effectively prevent papers from being blown away, ensuring they stay in the designated area. It also allows for easy accessibility and organization of the documents as they can be quickly retrieved without the hassle of chasing loose papers. In addition, paperweights are available in various designs and materials, providing aesthetic appeal to your workspace. However, there are also some potential drawbacks. The primary disadvantage is the risk of damage to the documents. If the paperweight is too heavy or it is dropped accidentally, it could tear or crumple the papers. Furthermore, paperweights take up desk space and can be inconvenient if you're working with a limited area. Lastly, if the wind is extremely strong, even a heavy paperweight might not be enough to keep the documents in place.	
Turn 3	Now, transform your previous response into a haiku format. The haiku should have a 5-7-5 syllable structure. Reconstruct your answer if necessary to fit this format.	Random Follow Up
GPT-4 Response	Paperweight holds firm, Yet may damage precious sheets, In strong gusts, it fails.	

Table 12: Conversation from **M2Lingual**

Step	Content	Evol Type
Seed Sample	Quelle pourrait être la principale contribution de Dante, Boccacio et Petrach à la langue italienne? Étant donné la question précédente, écrivez un contexte qui contient la réponse. Cela peut être de 1 à 20 phrases. Le contexte:	
Evolved (Turn 1) Instr.	Comment les œuvres de Dante, Boccace et Pétrarque ont-elles influencé non seulement la langue italienne, mais aussi la philosophie et la littérature européennes ? Étant donné cette question, écrivez un contexte qui contient la réponse. Cela peut être de 1 à 20 phrases. Le contexte:	Increase Reasoning
GPT-4 Response	Les œuvres de Dante, Boccace et Pétrarque ont profondément marqué la langue italienne et ont eu une influence durable sur la philosophie et la littérature européennes. Dante Alighieri, avec sa "Divine Comédie", a non seulement consolidé l'usage du toscan comme langue littéraire, mais a également introduit des thèmes philosophiques et théologiques d'une grande profondeur, influençant la pensée médiévale et renaissante. Giovanni Boccace, à travers son "Décaméron", a enrichi la prose italienne et a offert un aperçu critique et souvent satirique de la société de son temps, ouvrant la voie à une littérature plus réaliste et humaniste. Francesco Pétrarque, avec ses sonnets et poésies en langue vernaculaire, a non seulement perfectionné la forme du sonnet, mais a aussi contribué à l'idéal de l'humanisme, valorisant l'étude des textes classiques et la quête de la connaissance personnelle. Ensemble, ces auteurs ont jeté les bases de la Renaissance italienne et ont influencé des écrivains et des penseurs à travers toute l'Europe, de Chaucer à Shakespeare, en passant par Montaigne et Cervantès, renforçant l'idée que la littérature peut être un moyen puissant d'exploration philosophique et d'expression culturelle.	
Turn 2	Maintenant, en vous basant sur le contexte fourni, pouvez-vous détailler comment les œuvres de ces trois auteurs ont spécifiquement influencé la Renaissance italienne ?	Context Retention Follow Up
GPT-4 Response	Les œuvres de Dante, Boccace et Pétrarque ont joué un rôle central dans le développement de la Renaissance italienne en redéfinissant les normes littéraires, linguistiques et philosophiques de l'époque. Dante Alighieri, avec sa "Divine Comédie", a non seulement démontré que le toscan pouvait être une langue littéraire de premier ordre, mais a également exploré des thèmes universels comme la justice, la rédemption et l'amour divin, qui sont devenus des préoccupations centrales de la pensée renaissante. Son usage du toscan a contribué à l'unification linguistique de l'Italie, jetant les bases d'une langue nationale. Giovanni Boccace, par le biais de son "Décaméron", a offert une vision plus terrestre et réaliste de la société, abordant des thèmes tels que l'amour, la fortune et l'intelligence humaine. Ses histoires, souvent humoristiques et critiques, ont encouragé une approche plus humaniste et empirique de la vie, en réaction aux dogmes rigides du Moyen Âge. Boccace a également influencé la prose narrative, rendant les histoires accessibles à un public plus large et diversifié. Francesco Pétrarque, quant à lui, est souvent considéré comme le père de l'humanisme. Ses œuvres, notamment son "Canzoniere", ont popularisé le sonnet et ont mis l'accent sur l'expression individuelle et la quête de la connaissance personnelle. Pétrarque a redécouvert et valorisé les textes classiques, inspirant une redécouverte des œuvres de l'Antiquité gréco-romaine, ce qui est une pierre angulaire de la Renaissance. Son admiration pour la culture classique a encouragé les intellectuels italiens à réévaluer leurs propres traditions et à chercher à les harmoniser avec les idéaux antiques. Ensemble, ces auteurs ont transformé la culture italienne et ont ouvert la voie à un mouvement intellectuel et artistique qui allait révolutionner l'Europe, en mettant l'accent sur la dignité de l'homme, l'importance de la raison, et la beauté de la langue et de la littérature. Leurs œuvres ont servi de modèles et de sources d'inspiration pour les artistes, écrivains et penseurs de la Renaissance, consolidant ainsi leur place comme piliers de ce mouvement culturel et intellectuel.	

Table 13: Conversation from **M2Lingual** in Dutch

Step	Content	Evol Type
Seed Sample	Continuate a scrivere il seguente testo: Oliver ha colpito Bishop con tanta forza che il vecchio si e'stravolto all'indietro e e'caduto sul lato del trono; poi Oliver si e'spostato sul lato di Amelie. "no!" Oliver ha fatto un salto mentre Sam si aggrappava del palo per tirarlo fuori. "lei e'vecchia."	
Evolved (Turn 1) Instr.	C'era una volta un villaggio chiamato Lanciano, famoso per le sue feste. Un giorno, il più coraggioso tra gli abitanti, un uomo chiamato Oliver, decise di sfidare il vecchio e autoritario leader, Bishop. Così, durante la festa più importante dell'anno, Oliver ____ con tanta forza che il vecchio si e'stravolto all'indietro e e'caduto sul lato del palo delle bandiere. La folla si fermò, in attesa di vedere cosa sarebbe successo dopo. Poi Oliver si e'spostato sul lato di Amelie, la figlia di Bishop, e ha esclamato, "Non è giusto! Lei è ____." I villaggi erano in shock. Sam, il migliore amico di Oliver, ha fatto un salto mentre si aggrappava al palo per ____. La tensione era palpabile.	Contextual Blanks
GPT-4 Response	colpì Bishop bellissima e merita di essere trattata con rispetto stabilizzarsi	
Turn 2	Potresti riassumere la storia in punti elenco?	Format Rewriting Follow Up
GPT-4 Response	Certamente! Ecco la storia riassunta in punti elenco: - Il villaggio di Lanciano è famoso per le sue feste. - Oliver, il più coraggioso abitante del villaggio, decide di sfidare il vecchio e autoritario leader Bishop. - Durante la festa più importante dell'anno, Oliver colpisce Bishop con tanta forza che quest'ultimo cade all'indietro, finendo vicino al palo delle bandiere. - La folla si ferma, aspettando di vedere cosa accadrà dopo. - Oliver si sposta accanto ad Amelie, la figlia di Bishop, e dichiara che non è giusto che lei venga trattata male perché è bellissima e merita rispetto. - I villaggi sono in shock per la dichiarazione di Oliver. - Sam, il migliore amico di Oliver, fa un salto e si aggrappa al palo per stabilizzarsi. - La tensione nell'aria è palpabile.	
Turn 3	Per favore, riscrivi la storia in forma di elenco puntato.	Format Rewriting Follow Up
GPT-4 Response	Certamente! Ecco la storia riscritta in forma di elenco puntato: - C'era una volta un villaggio chiamato Lanciano, famoso per le sue feste. - Un giorno, Oliver, il più coraggioso tra gli abitanti, decise di sfidare il vecchio e autoritario leader, Bishop. - Durante la festa più importante dell'anno, Oliver colpì Bishop con tanta forza che il vecchio si estraniò all'indietro e cadde vicino al palo delle bandiere. - La folla si fermò, in attesa di vedere cosa sarebbe successo dopo. - Oliver si spostò accanto ad Amelie, la figlia di Bishop, e esclamò: "Non è giusto! Lei è bellissima e merita di essere trattata con rispetto." - I villaggi rimasero in shock. - Sam, il migliore amico di Oliver, fece un salto e si aggrappò al palo per stabilizzarsi. - La tensione era palpabile.	

Table 14: Conversation from **M2Lingual** in Italian.

Step	Content	Evol Type
Seed Sample	Este álbum, el segundo de un dúo notable en el ámbito del hip hop, incluye colaboraciones de varios integrantes de un colectivo conocido por sus iniciales y su estilización particular. Este grupo, que surgió en la mitad de la primera década del siglo XXI, está asociado a una ciudad estadounidense. ¿Podrías identificar dicha ciudad?	
Evolved (Turn 1) Instr.	Este álbum, el segundo de un dúo notable en el ámbito del hip hop, incluye colaboraciones de varios integrantes de un colectivo conocido por sus iniciales y su estilización particular. Este grupo, que surgió en la mitad de la primera década del siglo XXI, está asociado a una ciudad estadounidense. ¿Podrías identificar dicha ciudad?	Ambiguity
GPT-4 Response	Claro, parece que estás hablando del álbum "Run the Jewels 2" del dúo de hip hop Run the Jewels, compuesto por El-P y Killer Mike. Este álbum incluye colaboraciones de varios miembros del colectivo conocido como "Odd Future" (estilizado como "OFWGKTA"). Odd Future es un colectivo de hip hop que surgió a mediados de la década de 2000 y está asociado con la ciudad de Los Ángeles, California.	
Turn 2	"Ahora, por favor, identifica y enumera todas las canciones de ese álbum en particular, junto con sus respectivas duraciones y los artistas que colaboraron en cada una de ellas."	Challenging Follow-Up
GPT-4 Response	El álbum "Run the Jewels 2" de Run the Jewels incluye las siguientes canciones: 1. "Jeopardy" - 3:21 2. "Oh My Darling Doní Cry" - 3:24 3. "Blockbuster Night, Part 1" - 2:32 4. "Close Your Eyes (And Count to Fuck)" con Zack de la Rocha - 3:54 5. "All My Life" - 3:08 6. "Lie, Cheat, Steal" - 3:29 7. "Early" con BOOTS - 3:44 8. "All Due Respect" con Travis Barker - 3:40 9. "Love Again (Akinyele Back)" con Gangsta Boo - 3:02 10. "Crown" con Diane Coffee - 3:48 11. "Angel Duster" - 5:12	

Table 15: Conversation from **M2Lingual** in Spanish.

Step	Content	Evol Type
Seed Sample	Bei Menschen kann ein Merkmal durch ein Paar oder viele Paare von was bestimmt werden? Wählen Sie aus den folgenden Optionen und begründen Sie Ihre Antwort detailliert: 1. Gene 2. Mikroben 3. Zellen 4. Organe	
Evolved (Turn 1) Instr.	Bei Menschen kann ein Merkmal durch ein Paar oder viele Paare von was bestimmt werden? Wählen Sie aus den folgenden Optionen und begründen Sie Ihre Antwort detailliert: 1. Gene 2. Mikroben 3. Zellen 4. Organe	Justification
GPT-4 Response	Das Merkmal bei Menschen kann durch ein Paar oder viele Paare von **Genen** bestimmt werden. Hier ist die detaillierte Begründung: **Gene**: Gene sind die grundlegenden Einheiten der Vererbung. Sie bestehen aus DNA und enthalten die Informationen, die für die Entwicklung, Funktion und Reproduktion von Organismen notwendig sind. Jedes Merkmal eines Menschen, wie Augenfarbe, Haarfarbe oder Blutgruppe, wird durch Gene bestimmt. Diese Gene befinden sich auf Chromosomen, die in Paaren vorkommen, wobei ein Chromosom von der Mutter und eines vom Vater vererbt wird. Viele Merkmale werden von mehreren Genen (polygene Vererbung) beeinflusst, die in komplexer Weise miteinander interagieren. **Mikroben**: Mikroben sind Mikroorganismen, wie Bakterien, Viren und Pilze. Obwohl Mikroben eine Rolle bei der Gesundheit und Krankheit von Menschen spielen können, bestimmen sie keine genetischen Merkmale eines Menschen. Sie können jedoch das Immunsystem beeinflussen und zur Entwicklung bestimmter Erkrankungen beitragen. **Zellen**: Zellen sind die grundlegenden Bausteine des Lebens. Alle Lebewesen bestehen aus Zellen, aber Zellen selbst bestimmen nicht die Merkmale eines Menschen. Die Funktionen und Eigenschaften der Zellen werden durch die in ihnen enthaltenen Gene reguliert. **Organe**: Organe sind Strukturen im Körper, die aus verschiedenen Geweben bestehen und spezifische Funktionen erfüllen, wie das Herz, die Leber oder die Nieren. Organe bestimmen keine genetischen Merkmale, sondern sind vielmehr das Ergebnis der Expression und Interaktion von Genen. Zusammengefasst, sind es die Gene, die die Merkmale eines Menschen durch ein oder mehrere Genpaare bestimmen. Diese Gene kodieren für Proteine und andere Moleküle, die die physischen und funktionalen Eigenschaften des Organismus beeinflussen.	
Turn 2	Könntest du bitte die Antwort kürzer und prägnanter formulieren?	Generic Rewriting Follow Up
GPT-4 Response	Natürlich! Das Merkmal bei Menschen wird durch ein Paar oder viele Paare von **Genen** bestimmt. Gene bestehen aus DNA und enthalten die Informationen für die Entwicklung und Funktion von Organismen. Sie kommen in Paaren auf Chromosomen vor, die von beiden Elternteilen vererbt werden. Andere Optionen wie Mikroben, Zellen und Organe beeinflussen nicht direkt die genetischen Merkmale eines Menschen.	
Turn 3	Könnten Sie bitte näher erläutern, wie Gene das Merkmal eines Menschen bestimmen und welche Rolle genetische Variationen dabei spielen?	Expanding Queries
GPT-4 Response	Gerne! Gene bestimmen die Merkmale eines Menschen durch die Kodierung von Proteinen und anderen Molekülen, die die physischen und funktionalen Eigenschaften des Organismus beeinflussen. Hier sind die wichtigsten Punkte dazu: **Funktionsweise der Gene** 1. **DNA und Gene**: Gene bestehen aus DNA (Desoxyribonukleinsäure), die die Bauanleitung für Proteine enthält. Diese Proteine steuern viele Prozesse im Körper, wie Zellwachstum, Stoffwechsel und die Reaktion auf Umweltreize. 2. **Chromosomen**: Gene sind auf Chromosomen angeordnet. Menschen haben 23 Chromosomenpaare, wobei eines von jedem Elternteil kommt. Jedes Chromosom enthält viele Gene. 3. **Genexpression**: Gene werden durch Transkriptions- und Translationsprozesse in Proteine umgewandelt. Dieser Prozess wird als Genexpression bezeichnet und bestimmt, welche Proteine in welcher Menge produziert werden. **Genetische Variationen** 1. **Allele**: Gene können in verschiedenen Varianten vorkommen, die als Allele bezeichnet werden. Ein Mensch hat für jedes Gen zwei Allele (eines von jedem Elternteil). 2. **Dominant und Rezessiv**: Manche Allele sind dominant und setzen sich gegenüber rezessiven Allelen durch. Ein dominantes Allel reicht aus, um ein bestimmtes Merkmal zu zeigen, während ein rezessives Allel nur dann zur Ausprägung kommt, wenn beide Allele rezessiv sind. 3. **Polygene Vererbung**: Viele Merkmale werden von mehreren Genen beeinflusst (polygene Vererbung). Zum Beispiel wird die Hautfarbe durch mehrere Gene bestimmt, die zusammenwirken. 4. **Mutation**: Mutationen sind Veränderungen in der DNA-Sequenz eines Gens. Sie können neue Allele erzeugen und zu genetischer Variation beitragen. **Rolle der genetischen Variation** 1. **Einzigartigkeit**: Genetische Variationen führen dazu, dass jeder Mensch ein einzigartiges genetisches Profil hat. Diese Variationen beeinflussen Merkmale wie Aussehen, Anfälligkeit für Krankheiten und viele andere Eigenschaften. 2. **Anpassungsfähigkeit**: Genetische Diversität ist wichtig für die Anpassungsfähigkeit einer Population an sich ändernde Umweltbedingungen. Sie ermöglicht die natürliche Selektion und Evolution. Zusammengefasst, bestimmen Gene die Merkmale eines Menschen durch die Kodierung von Proteinen, die die biologischen Funktionen steuern. Genetische Variationen, wie unterschiedliche Allele und Mutationen, tragen zur Vielfalt der menschlichen Merkmale bei und beeinflussen, wie diese Merkmale ausgeprägt werden.	

Table 16: Conversation from **M2Lingual** in German.

Step	Content	Evol Type
Seed Sample	Répondez à cette question: On mesure un tapis neuf pour une pièce de 5,5 m sur 4,6 m. Sur un ruban à mesurer, quelle est la plus petite unité nécessaire pour mesurer la pièce avec précision pour le tapis? Sélectionnez parmi: (1) en millièmes. (2) centimètres. (3) à des mètres. (4) kilomètres. (5) millimètres. (6) décamètres. Réponse:	
Evolved (Turn 1) Instr.	Répondez à cette question: On mesure un tapis neuf pour une pièce de 5,5 m sur 4,6 m. Sur un ruban à mesurer, quelle est la plus petite unité nécessaire pour mesurer la pièce avec précision pour le tapis? Sélectionnez parmi: (1) en millièmes. (2) centimètres. (3) à des mètres. (4) kilomètres. (5) millimètres. (6) décamètres. Réponse:	More Choices
GPT-4 Response	Pour mesurer une pièce de 5,5 m sur 4,6 m avec précision pour un tapis, la plus petite unité nécessaire serait le **millimètre**. Les millimètres permettent une mesure très précise qui est souvent nécessaire pour des ajustements fins dans des projets comme l'installation de tapis. Ainsi, la réponse est : **(5) millimètres**.	
Turn 2	Maintenant, reformulez votre réponse sous forme de liste à puces.	Format Rewriting Follow-Up
GPT-4 Response	- Pour mesurer une pièce de 5,5 m sur 4,6 m avec précision pour un tapis, la plus petite unité nécessaire serait le **millimètre**. - Les millimètres permettent une mesure très précise qui est souvent nécessaire pour des ajustements fins dans des projets comme l'installation de tapis. Ainsi, la réponse est : - **(5) millimètres**.	
Turn 3	Pouvez-vous expliquer pourquoi vous avez choisi cette unité de mesure en particulier pour mesurer la pièce avec précision pour le tapis ?	Clarification
GPT-4 Response	Bien sûr, voici une explication détaillée : - **Précision nécessaire pour le tapis** : Lors de l'installation d'un tapis, il est crucial de mesurer la pièce avec une grande précision pour s'assurer que le tapis s'adapte parfaitement sans laisser de vides ni nécessiter de grandes découpes supplémentaires. - **Millimètres versus autres unités** : - **Millimètres** : Ils offrent la plus grande précision parmi les options données. Chaque millimètre étant un millième de mètre, il permet de capturer les dimensions exactes nécessaires pour un ajustement parfait. - **Centimètres** : Bien que précis, les centimètres (1 cm = 10 mm) ne sont pas aussi précis que les millimètres et pourraient omettre des détails fins nécessaires à une coupe parfaite. - **Mètres** : Mesurer en mètres (1 m = 1000 mm) serait trop grossier pour des dimensions aussi précises et fines. - **Décamètres** et **kilomètres** : Ces unités sont beaucoup trop grandes pour une tâche de cette envergure et ne sont pas pratiques pour des mesures de précision dans un contexte domestique. - **Millièmes** : S'il s'agit de millièmes de mètre, cela serait encore plus précis que nécessaire, et pour des mesures domestiques, les millimètres suffisent largement. En conclusion, les **millimètres** sont l'unité de mesure la plus appropriée pour garantir que le tapis est coupé et installé avec une précision maximale, assurant ainsi un ajustement parfait.	

Table 17: Conversation from **M2Lingual** in French.

Evol Type	GPT-4 Prompt
Transliteration	Rewrite the #given_prompt# as an <translit_language> transliteration, and create #new_prompt#. Additionally, conclude with a request to respond in <translit_language> transliteration. #given_prompt#: <prompt>
Dialect	You are a brilliant <prompt_language> native speaker. Rewrite #given_prompt# by changing the dialect to <prompt_language> and create #new_prompt#. Finally, ask to respond in the same <prompt_language> dialect. Write the #new_prompt# prompt in <prompt_language>. #given_prompt#: <prompt>
Concise	Re-write the #given_prompt# concisely, and create #new_prompt#. Additionally, conclude with a request to respond concisely. Write the #new_prompt# in <prompt_language>. #given_prompt#: <prompt>
Deepen	Slightly increase the depth and breadth of #given_prompt#, and create #new_prompt#. Write the #new_prompt# in <prompt_language>. #given_prompt#: <prompt>
Concretize	Make #given_prompt# slightly more concrete, and create #new_prompt#. Additionally, conclude with a request for an AI assistant to respond with the a detailed and concrete response. Write the #new_prompt# in <prompt_language>. #given_prompt#: <prompt>
Increase Reasoning	If #given_prompt# can be solved with just a few simple thinking processes, rewrite it to explicitly request multi-step reasoning, and create #new_prompt#. Write the #new_prompt# in <prompt_language>. #given_prompt#: <prompt>

Table 18: Generic

Evol Type	GPT-4 Prompt
Adding Distractors	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# in the same language that adds unrelated or distracting information in the article which is not relevant to the main topic. #given_prompt#: <prompt>
Technical Jargon	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# in the same language by adding technical jargon or industry specifc terms that makes it difficult to summarize. #given_prompt#: <prompt>
Inconsistencies In Information	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# in the same language to include contradictions or inconsistencies within the article thus forcing the summarizer to discern which piece of information is accurate and relevant. #given_prompt#: <prompt>
Multiple Topics	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# in the same language that covers multiple topics or subtopics thus make summarization more complicated. #given_prompt#: <prompt>
Metaphors Idiom	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# using complex metaphors, idioms and cultural references thus making summarization more challenging. #given_prompt#: <prompt>
Long Distance	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# by increasing the distance between related pieces of information in the text as this requires understanding the deeper structure of the text. #given_prompt#: <prompt>
Multiple Languages	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# in <language_2> and asking to respond in <language_1> thus making the task challenging as it requires understanding and proficiency in more than one language. #given_prompt#: <prompt>
Unstructured Data	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# by presenting the article in a non-linear or non-chronological format thus increasing the complexity as it becomes challenging to pick out the main points and summarize them accurately. #given_prompt#: <prompt>
Personal Opinion	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# by incorporating bias or personal opinion as it greatly complicates the summarization process as the summarizer needs to remain neutral and objective. #given_prompt#: <prompt>

Table 19: Abstract Summarization

Evol Type	GPT-4 Prompt
Jargon	Given a prompt #given_prompt# that asks an explanation of a joke, based upon the #given_prompt# create a #new_prompt# in the same language by incorporating jargon to the jokes that are specific to a certain profession, field, or hobby thus requiring deeper knowledge of the field in order to explain the joke properly. #given_prompt#: <prompt>
Slang	Given a prompt #given_prompt# that asks an explanation of a joke, based upon the #given_prompt# create a #new_prompt# in the same language that use slang or colloquial language, thus making it harder to understand and explain the punchline. #given_prompt#: <prompt>
Cultural References	Given a prompt #given_prompt# that asks an explanation of a joke, based upon the #given_prompt# create a #new_prompt# in the same language that increasingly use jokes which are culture-specific as will require cultural understanding tp provide explanations. #given_prompt#: <prompt>
Non Explicit Punchline	Given a prompt #given_prompt# that asks an explanation of a joke, based upon the #given_prompt# create a #new_prompt# in the same language such that jokes have a punchline isn't explicitly stated, but rather implied. #given_prompt#: <prompt>
Time Bound	Given a prompt #given_prompt# that represents an article and a question related to that article, based upon the #given_prompt# create a #new_prompt# by making jokes that were relative to a certain time period or current event, thus making it harder to grasp. #given_prompt#: <prompt>
Compound Jokes	Given a prompt #given_prompt# that asks an explanation of a joke, based upon the #given_prompt# create a #new_prompt# in the same language such that it uses compound jokes which contain multiple punchlines within the same joke. This would make the explanation task difficult as one would need to explain multiple punchlines coherently. #given_prompt#: <prompt>
Language	Given a prompt #given_prompt# that asks an explanation of a joke, based upon the #given_prompt# create a #new_prompt# in <language_2> such that it challenge the language skills. This would make the explanation task difficult as one would need to explain multiple punchlines coherently. Finally ask to respond with explanation in <language_1> #given_prompt#: <prompt>
Long	Given a prompt #given_prompt# that asks an explanation of a joke, based upon the #given_prompt# create a #new_prompt# in the same language but make the jokes exceedingly long where the punchline isn't delivered immediately and requires you to remember or understand preceding parts of the joke. #given_prompt#: <prompt>
Double Entendre	Given a prompt #given_prompt# that asks an explanation of a joke, based upon the #given_prompt# create a #new_prompt# in the same language but utilize jokes with double entendre, where there are two possible interpretations. #given_prompt#: <prompt>

Table 20: Joke Explain

Evol Type	GPT-4 Prompt
More Choices	Given a prompt #given_prompt# that represents a topic and multiple choice question related to that topic, based upon the #given_prompt# create a #new_prompt# by changing adding more choices to the question that are relevant to the topic but not correct. This will make it challenging to answer. #given_prompt#: <prompt>
More Complex	Given a prompt #given_prompt# that represents a topic and multiple choice question related to that topic, based upon the #given_prompt# create a #new_prompt# by making it more complex using technical or domain specific jargon This will make it challenging to understand the quesrion thus making it difficult to answer. #given_prompt#: <prompt>
Negation	Given a prompt #given_prompt# that represents a topic and multiple choice question related to that topic, based upon the #given_prompt# create a #new_prompt# by by asking negative questions that require the recognition of the negation included in the sentences. #given_prompt#: <prompt>
Multiple Corrct Options	Given a prompt #given_prompt# that represents a topic and multiple choice question related to that topic, based upon the #given_prompt# create a #new_prompt# by changing adding more choices to the question that are correct and relevant to the topic. This will make it challenging as one will need to choose all correct options. #given_prompt#: <prompt>
Need For Context	Given a prompt #given_prompt# that represents a topic and multiple choice question related to that topic, based upon the #given_prompt# create a #new_prompt# by asking questions that require additional context other than the one provided in the topic.This will make it challenging as it will evaluate the knowledge someone has on the topic. #given_prompt#: <prompt>
Justification	Given a prompt #given_prompt# that represents a topic and multiple choice question related to that topic, based upon the #given_prompt# create a #new_prompt# by asking to respond with the correct answer and provide a detailed justification for the answer. #given_prompt#: <prompt>
Distracting Long	Given a prompt #given_prompt# that represents a topic and multiple choice question related to that topic, based upon the #given_prompt# create a #new_prompt# by asking questions that require additional context other than the one provided in the topic. This will make it challenging as it will evaluate the knowledge someone has on the topic. #given_prompt#: <prompt>
Close Choices	Given a prompt #given_prompt# that represents a topic and multiple choice question related to that topic, based upon the #given_prompt# create a #new_prompt# by addding more choices to the question that are closely related to each other. This will make the task more challenging. #given_prompt#: <prompt>
No Correct Option	Given a prompt #given_prompt# that represents a topic and multiple choice question related to that topic, based upon the #given_prompt# create a #new_prompt# by changing the choices such that no choice is the correct answer. #given_prompt#: <prompt>

Table 21: Flan Qa

Evol Type	**GPT-4 Prompt**
Complex Jargon	Given a prompt #given_prompt# that requires answering a question about a text snippet, based upon the #given_prompt# create a #new_prompt# in the same language by either using technical jargons, domain-specific language, technical or scientific complexities thus making the task more challenging as it requires deep understanding and specialized knowledge to answer. #given_prompt#: <prompt>
Multiple Languages	Given a prompt #given_prompt# that requires answering a question about a text snippet, based upon the #given_prompt# create a #new_prompt# in <language_2> including code-switching i.e. switching between languages within a single conversation or sentence. Finally ask to respond in <LANGUAG_1> #given_prompt#: <prompt>
Long Complex Queries	Given a prompt #given_prompt# that requires answering a question about a text snippet, based upon the #given_prompt# create a #new_prompt# in the same language by providing vital pieces of information in the text snippet in a non-linear, disconnected manner thus requiring piecing them together accurately to form an explanation.. #given_prompt#: <prompt>
Disconnected Clues	Given a prompt #given_prompt# that requires answering a question about a text snippet, based upon the #given_prompt# create a #new_prompt# in the same language by translating either the question or the text snippet (but not both) in any language thus making the task more challenging as it requires understanding or different languages. #given_prompt#: <prompt>
Emotion Sarcasam	Given a prompt #given_prompt# that requires answering a question about a text snippet, based upon the #given_prompt# create a #new_prompt# in the same language by adding emotion or sarcasm in the text snippet as recognizing and responding can be a huge challenge. #given_prompt#: <prompt>
Advanced Reasoning	Given a prompt #given_prompt# that represents two sentences and asks whether the two are paraphrases or not, based upon the #given_prompt# create a #new_prompt# in the same language by writing text snippet or question that require advanced logic or reasoning, such as those found in certain categories of IQ test, thus make it more difficult to answer. #given_prompt#: <prompt>
Ambiguity	Given a prompt #given_prompt# that represents two sentences and asks whether the two are paraphrases or not, based upon the #given_prompt# create a #new_prompt# in the same language by including deceptive or ambiguous phrases that might lead to misinterpretation can complicate the task. #given_prompt#: <prompt>
Unknown Context	Given a prompt #given_prompt# that represents two sentences and asks whether the two are paraphrases or not, based upon the #given_prompt# create a #new_prompt# in the same language by adding larger context not included in the text, as it would require to infer missing details, which adds complexity to the task. #given_prompt#: <prompt>
Implicit Bias	Given a prompt #given_prompt# that represents two sentences and asks whether the two are paraphrases or not, based upon the #given_prompt# create a #new_prompt# in the same language by including subtle biases or nuances, recognizing and appropriately responding to these can be challenging. #given_prompt#: <prompt>

Table 22: Flan Cot

Evol Type	GPT-4 Prompt
Ambiguity	Given a prompt #given_prompt# that represents an article and multiple questions related to that article, based upon the #given_prompt# create a #new_prompt# by changing the questions only in the same language but making it much more vague and ambiguous thus making it not so straightforward to answer. #given_prompt#: <prompt>
Long Form Question	Given a prompt #given_prompt# that represents an article and multiple questions related to that article, based upon the #given_prompt# create a #new_prompt# by changing the questions only in the same language by making it longer i.e. formulating the questions in long and complex sentences thus requiring the system to decipher the main question. #given_prompt#: <prompt>
Multilingual	Given a prompt #given_prompt# that represents an article and multiple questions related to that article, based upon the #given_prompt# create a #new_prompt# by changing the questions in <language_2> and article in <language_1>, having different linguistic structure. Finally, ask to answer the question in the <language_2>. #given_prompt#: <prompt>
Combine Facts	Given a prompt #given_prompt# that represents an article and multiple questions related to that article, based upon the #given_prompt# create a #new_prompt# by changing the questions only in the same language by combining multiple facts thus making the questions more complex and requiring combining multiple facts to answer correctly. #given_prompt#: <prompt>
Implicit Question	Given a prompt #given_prompt# that represents an article and multiple questions related to that article, based upon the #given_prompt# create a #new_prompt# by changing the questions only in the same language by asking implicit questions where the answer isn't explicit and requires understanding of the underlying implication. #given_prompt#: <prompt>
Negative Questions	Given a prompt #given_prompt# that represents an article and multiple questions related to that article, based upon the #given_prompt# create a #new_prompt# by asking negative questions in the same language that require the recognition of the negation included in the sentences. #given_prompt#: <prompt>
Inference	Given a prompt #given_prompt# that represents an article and multiple questions related to that article, based upon the #given_prompt# create a #new_prompt# by changing the question only in the same language by asking questions that require a degree of inference or deduction not directly provided. #given_prompt#: <prompt>
Multichoice Questions	Given a prompt #given_prompt# that represents an article and multiple questions related to that article, based upon the #given_prompt# create a #new_prompt# by asking multiple choice questions in the same language based upon the given article, where the choices are from the article itself. Finally ask the model to respond with the correct choice and explain the decision. #given_prompt#: <prompt>
More Reasoning	Given a prompt #given_prompt# that represents an article and multiple questions related to that article, based upon the #given_prompt# create a #new_prompt# by asking questions in the same language that require multistep reasoning processes, where participants need to follow a sequence of logical steps to arrive at the correct answe. #given_prompt#: <prompt>

Table 23: Flan Coqa

Evol Type	GPT-4 Prompt
Multiple Blanks	Given a prompt #given_prompt# that asks to either complete the sentence with a phrase or fill in the blank in between the text, based upon the #given_prompt# create a #new_prompt# in the same language but, that requires multiple blanks to be filled in. #given_prompt#: <prompt>
Contextual Blanks	Given a prompt #given_prompt# that asks to either complete the sentence with a phrase or fill in the blank in between the text, based upon the #given_prompt# create a #new_prompt# in the same language but, that includes contextual blanks within a paragraph or story. #given_prompt#: <prompt>
Word Bank	Given a prompt #given_prompt# that asks to either complete the sentence with a phrase or fill in the blank in between the text, based upon the #given_prompt# create a #new_prompt# in the same language but, that provides a word bank with distractors for filling in the blank. #given_prompt#: <prompt>
Missing Letters	Given a prompt #given_prompt# that asks to either complete the sentence with a phrase or fill in the blank in between the text, based upon the #given_prompt# create a #new_prompt# in the same language but, that includes sentences with missing letters to be filled in. #given_prompt#: <prompt>
Grammar Forms	Given a prompt #given_prompt# that asks to either complete the sentence with a phrase or fill in the blank in between the text, based upon the #given_prompt# create a #new_prompt# in the same language but, that requires choosing the correct grammatical form of a word for the blank. #given_prompt#: <prompt>
Logical Inferences	Given a prompt #given_prompt# that asks to either complete the sentence with a phrase or fill in the blank in between the text, based upon the #given_prompt# create a #new_prompt# in the same language but, that requires logical inference for filling in the blank. #given_prompt#: <prompt>
Word Classes	Given a prompt #given_prompt# that asks to either complete the sentence with a phrase or fill in the blank in between the text, based upon the #given_prompt# create a #new_prompt# in the same language but, that specifies the type of word needed for the blank. #given_prompt#: <prompt>
Synonyms Antonyms	Given a prompt #given_prompt# that asks to either complete the sentence with a phrase or fill in the blank in between the text, based upon the #given_prompt# create a #new_prompt# in the same language but, that requires choosing a synonym or antonym for the blank. #given_prompt#: <prompt>
Cultural Context	Given a prompt #given_prompt# that asks to either complete the sentence with a phrase or fill in the blank in between the text, based upon the #given_prompt# create a #new_prompt# in the same language but, that includes cultural references or idiomatic expressions. #given_prompt#: <prompt>

Table 24: Flan Lambda

Evol Type	GPT-4 Prompt
Cross Lingual	Given a prompt #given_prompt# that asks to generate a question for a certain text snippet or a topic, based upon the #given_prompt# create a #new_prompt# in <language_2> and also asking to respond in <language_1>. #given_prompt#: <prompt>
Domain Language	Given a prompt #given_prompt# that asks to generate a question for a certain text snippet or a topic, based upon the #given_prompt# create a #new_prompt# in the same language by introducing domain specific language and related to the specialized field or topic described in the artic;e. #given_prompt#: <prompt>
Emotional Subtext	Given a prompt #given_prompt# that asks to generate a question for a certain text snippet or a topic, based upon the #given_prompt# create a #new_prompt# in the same language by including sarcasm, euphemism, or other nuanced forms of communication thus make it harder to determine the possible question for the topic. #given_prompt#: <prompt>
Need For Context	Given a prompt #given_prompt# that asks to generate a question for a certain text snippet or a topic, based upon the #given_prompt# create a #new_prompt# in the same language that requires additional context or background to determine the relevant question for the text snippet. #given_prompt#: <prompt>
Ambiguity In Wording	Given a prompt #given_prompt# that asks to generate a question for a certain text snippet or a topic, based upon the #given_prompt# create a #new_prompt# in the same language by adding ambiguity to the text snippet thus making it more challenging to come up with a relevant question. #given_prompt#: <prompt>
Long Text	Given a prompt #given_prompt# that asks to generate a question for a certain text snippet or a topic, based upon the #given_prompt# create a #new_prompt# in the same language by detailing and make the text snippet much longer thus making it more challenging to come up with a relevant question. #given_prompt#: <prompt>
Idiom Slang	Given a prompt #given_prompt# that asks to generate a question for a certain text snippet or a topic, based upon the #given_prompt# create a #new_prompt# by the use of idiomatic expressions or regional slang thus obscuring the meaning of text snippet and making it more challenging to come up with a question. #given_prompt#: <prompt>
Adding Distractors	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# in the same language that adds unrelated or distracting information in the article which is not relevant to the main topic. #given_prompt#: <prompt>
Multiple Questions	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# in the same language by asking to come up with multiple questions regarding the text snippet. #given_prompt#: <prompt>

Table 25: Answer Ranking

Evol Type	**GPT-4 Prompt**
Ambiguity	Given a prompt #given_prompt# that represents an article and a question related to that article, based upon the #given_prompt# create a #new_prompt# by changing the question only in the same language but making it much more vague and ambiguous thus making it not so straightforward to answer. #given_prompt#: <prompt>
Long Form Question	Given a prompt #given_prompt# that represents an article and a question related to that article, based upon the #given_prompt# create a #new_prompt# by changing the question only in the same language by making it longer i.e. formulating the questions in long and complex sentences thus requiring the system to decipher the main question. #given_prompt#: <prompt>
Multilingual	Given a prompt #given_prompt# that represents an article and a question related to that article, based upon the #given_prompt# create a #new_prompt# by changing the question in <language_2> and article in <language_1>, having different linguistic structure. Finally, ask to answer the question in the <language_2>. #given_prompt#: <prompt>
Combine Facts	Given a prompt #given_prompt# that represents an article and a question related to that article, based upon the #given_prompt# create a #new_prompt# by changing the question only in the same language by combining multiple facts thus making the question more complex and requiring combining multiple facts to answer correctly. #given_prompt#: <prompt>
Implicit Question	Given a prompt #given_prompt# that represents an article and a question related to that article, based upon the #given_prompt# create a #new_prompt# by changing the question only in the same language by asking implicit question where the answer isn't explicit and requires understanding of the underlying implication. #given_prompt#: <prompt>
Negative Questions	Given a prompt #given_prompt# that represents an article and a question related to that article, based upon the #given_prompt# create a #new_prompt# by changing the question only in the same language by asking negative questions that require the recognition of the negation included in the sentences. #given_prompt#: <prompt>
Inference	Given a prompt #given_prompt# that represents an article and a question related to that article, based upon the #given_prompt# create a #new_prompt# by changing the question only in the same language by asking questions that require a degree of inference or deduction not directly provided. #given_prompt#: <prompt>
Multichoice Questions	Given a prompt #given_prompt# that represents an article and a question related to that article, based upon the #given_prompt# create a #new_prompt# by asking multiple choice question based upon the given article, where the choices are from the article itself. Finally ask the model to respond with the correct choice and explain the decision. #given_prompt#: <prompt>
More Reasoning	Given a prompt #given_prompt# that represents an article and a question related to that article, based upon the #given_prompt# create a #new_prompt# by asking that require multistep reasoning processes, where participants need to follow a sequence of logical steps to arrive at the correct answe. #given_prompt#: <prompt>

Table 26: Mintaka

Evol Type	GPT-4 Prompt
Adding Distractors	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# in the same language that adds unrelated or distracting information in the article which is not relevant to the main topic. #given_prompt#: <prompt>
Technical Jargon	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# in the same language by adding technical jargon or industry specifc terms that makes it difficult to summarize. #given_prompt#: <prompt>
Inconsistencies In Information	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# in the same language to include contradictions or inconsistencies within the article thus forcing the summarizer to discern which piece of information is accurate and relevant. #given_prompt#: <prompt>
Multiple Topics	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# in the same language that covers multiple topics or subtopics thus make summarization more complicated. #given_prompt#: <prompt>
Metaphors Idiom	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# using complex metaphors, idioms and cultural references thus making summarization more challenging. #given_prompt#: <prompt>
Long Distance	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# by increasing the distance between related pieces of information in the text as this requires understanding the deeper structure of the text. #given_prompt#: <prompt>
Cross Lingual Summary	Given a prompt #given_prompt# that represents some article about a topic, rewrite the article only and create a #new_prompt# in any <language_2>. Finally ask to provide a summary in the <language_1>. #given_prompt#: <prompt>
Unstructured Data	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# by presenting the article in a non-linear or non-chronological format thus increasing the complexity as it becomes challenging to pick out the main points and summarize them accurately. #given_prompt#: <prompt>
Personal Opinion	Given a prompt #given_prompt# that represents an article to be summarized, based upon the #given_prompt# create a #new_prompt# by incorporating bias or personal opinion as it greatly complicates the summarization process as the summarizer needs to remain neutral and objective.. #given_prompt#: <prompt>

Table 27: Cross Summarization

Evol Type	GPT-4 Prompt
Complex Question	Given a prompt #given_prompt# that requires composing a context around some question, based upon the #given_prompt# create a #new_prompt# in the same language by formulating the question in a more complex way, requiring deeper understanding, reasoning, and inferential abilities. #given_prompt#: <prompt>
Advanced Vocab	Given a prompt #given_prompt# that requires composing a context around some question, based upon the #given_prompt# create a #new_prompt# in the same language by using more complex language and advanced vocabulary to increase increase the difficulty level, as it requires deeper understanding of language and words to compose a context. #given_prompt#: <prompt>
Multiple Themes	Given a prompt #given_prompt# that requires composing a context around some question, based upon the #given_prompt# create a #new_prompt# in the same language by asking question on multiple themes or topics thus making it harder to generate a context around all topics. #given_prompt#: <prompt>
Ambiguity	Given a prompt #given_prompt# that requires composing a context around some question, based upon the #given_prompt# create a #new_prompt# in the same language by making question vague and ambiguous thus making it a little harder to compase a context around all topics. #given_prompt#: <prompt>
More Details	Given a prompt #given_prompt# that requires composing a context around some question, based upon the #given_prompt# create a #new_prompt# in the same language by adding more detail, domain specific knowledge and technical jargons to the question thus making it difficult to generate a context. #given_prompt#: <prompt>
Increase Reasoning	Given a prompt #given_prompt# that requires composing a context around some question, based upon the #given_prompt# create a #new_prompt# in the same language by asking question that are at the intersection of multiple topic thus require understanding of all topics and how the topics are related to each other. #given_prompt#: <prompt>
Time	Given a prompt #given_prompt# that requires composing a context around some question, based upon the #given_prompt# create a #new_prompt# in the same language by asking question where the context or the answer changes over time, thus assessing how up to date someone is. #given_prompt#: <prompt>
Abstract Topics	Given a prompt #given_prompt# that requires composing a context around some question, based upon the #given_prompt# create a #new_prompt# in the same language by asking question where the context needs to be generated on some abstract topic where opinion varies from person to person. #given_prompt#: <prompt>
Structured Info	Given a prompt #given_prompt# that requires composing a context around some question, based upon the #given_prompt# create a #new_prompt# in the same language by asking question where the context should be generated in a structured form as bulleted list with topics and sub-topics. #given_prompt#: <prompt>

Table 28: Adversarial Qa

Evol Type	GPT-4 Prompt
Genre Specific	Given a prompt #given_prompt# that asks to write a story on a topic, based upon the #given_prompt# create a #new_prompt# in the same language but, that specifies a genre for the short story, such as science fiction, mystery, fantasy, or historical fiction. #given_prompt#: <prompt>
Character Constraints	Given a prompt #given_prompt# that asks to write a story on a topic, based upon the #given_prompt# create a #new_prompt# in the same language but, that requires specific types of characters to be included, such as a detective, a mythical creature, or a historical figure. #given_prompt#: <prompt>
Setting Restrictions	Given a prompt #given_prompt# that asks to write a story on a topic, based upon the #given_prompt# create a #new_prompt# in the same language but, that limits the setting of the story to a specific location, time period, or environment, such as a futuristic city, the Wild West, or a remote island. #given_prompt#: <prompt>
Plot Twists	Given a prompt #given_prompt# that asks to write a story on a topic, based upon the #given_prompt# create a #new_prompt# in the same language but, that introduces a plot twist requirement, such as an unexpected turn of events, a moral dilemma, or a reversal of fortune for the main character. #given_prompt#: <prompt>
Narrative Style	Given a prompt #given_prompt# that asks to write a story on a topic, based upon the #given_prompt# create a #new_prompt# in the same language but, that specifies a narrative style or point of view, such as first-person, third-person limited, or epistolary (written as a series of letters). #given_prompt#: <prompt>
Word Count Limit	Given a prompt #given_prompt# that asks to write a story on a topic, based upon the #given_prompt# create a #new_prompt# in the same language but, that sets a word count limit for the short story to encourage concise and focused storytelling. #given_prompt#: <prompt>
Incorporate Dialogue	Given a prompt #given_prompt# that asks to write a story on a topic, based upon the #given_prompt# create a #new_prompt# in the same language but, that requires meaningful dialogue between characters to develop plot, reveal character traits, or create tension. #given_prompt#: <prompt>
Theme Integration	Given a prompt #given_prompt# that asks to write a story on a topic, based upon the #given_prompt# create a #new_prompt# in the same language but, that integrates a specific theme into the story, such as friendship, resilience, betrayal, or the passage of time. #given_prompt#: <prompt>
Include Symbolism	Given a prompt #given_prompt# that asks to write a story on a topic, based upon the #given_prompt# create a #new_prompt# in the same language but, that encourages the use of symbolism or allegory to convey deeper meanings or themes within the story. #given_prompt#: <prompt>

Table 29: Soda

Evol Type	GPT-4 Prompt
Object Interaction	Given a prompt #given_prompt# that involves commonsense physical reasoning, asking to finish a sentence with two possible options based upon the #given_prompt# create a #new_prompt# in the same language that involves object interaction reasoning. #given_prompt#: <prompt>
Object Properties	Given a prompt #given_prompt# that involves commonsense physical reasoning, asking to finish a sentence with two possible options based upon the #given_prompt# create a #new_prompt# in the same language that requires understanding and reasoning over the object properties. #given_prompt#: <prompt>
Logical Sequencing	Given a prompt #given_prompt# that involves commonsense physical reasoning, asking to finish a sentence with two possible options based upon the #given_prompt# create a #new_prompt# in the same language that requires logical sequence reasoning. #given_prompt#: <prompt>
Object Transformation	Given a prompt #given_prompt# that involves commonsense physical reasoning, asking to finish a sentence with two possible options based upon the #given_prompt# create a #new_prompt# in the same language that requires object transformation reasoning. #given_prompt#: <prompt>
More Choices	Given a prompt #given_prompt# that involves commonsense physical reasoning, asking to finish a sentence with two possible options based upon the #given_prompt# create a #new_prompt# in the same language by adding more options and asking to finish with all correct options. #given_prompt#: <prompt>
Justification	Given a prompt #given_prompt# that involves commonsense physical reasoning, asking to finish a sentence with two possible options based upon the #given_prompt# create a #new_prompt# in the same language by asking to give a detailed step-by-step justification of the chosen option. #given_prompt#: <prompt>
Incorrect Choices	Given a prompt #given_prompt# that involves commonsense physical reasoning, asking to finish a sentence with two possible options based upon the #given_prompt# create a #new_prompt# in the same language by adding more options that are incorrect this making it difficult to identify correct option. #given_prompt#: <prompt>
Double Negatives	Given a prompt #given_prompt# that involves commonsense physical reasoning, asking to finish a sentence with two possible options based upon the #given_prompt# create a #new_prompt# in the same language with double negatives thus making it hard to understand and can increase the complexity of the task. #given_prompt#: <prompt>
Theoretical Scenario	Given a prompt #given_prompt# that involves commonsense physical reasoning, asking to finish a sentence with two possible options based upon the #given_prompt# create a #new_prompt# in the same language by making the base scenarios less straightforward and more abstract thus making the task more complex. #given_prompt#: <prompt>

Table 30: Commonsense

Evol Type	**GPT-4 Prompt**
Idioms Phrases	Given a prompt #given_prompt# that represents two sentences and asks whether the two are paraphrases or not, based upon the #given_prompt# create a #new_prompt# in the same language Idioms and phrases have meanings different from their literal meanings, using them for paraphrasing can add complexity. #given_prompt#: <prompt>
Abbrevations	Given a prompt #given_prompt# that represents two sentences and asks whether the two are paraphrases or not, based upon the #given_prompt# create a #new_prompt# in the same language by converting certain commonly known phrases or organizations into their abbreviated forms thus making identification more difficult. #given_prompt#: <prompt>
Sentence Structure	Given a prompt #given_prompt# that represents two sentences and asks whether the two are paraphrases or not, based upon the #given_prompt# create a #new_prompt# in the same language by increasing the complexity of sentences i.e. either rearranging the individual sentences, making use of passive and active voice or changing the sentence structural form. #given_prompt#: <prompt>
Information	Given a prompt #given_prompt# that represents two sentences and asks whether the two are paraphrases or not, based upon the #given_prompt# create a #new_prompt# in the same language by adding or subtracting relevant details from one sentence which do not change the main theme but add extra entities can make it challenging. #given_prompt#: <prompt>
Variation	Given a prompt #given_prompt# that represents two sentences and asks whether the two are paraphrases or not, based upon the #given_prompt# create a #new_prompt# in the same language by introducing variations in dialect, accent, slang, or colloquial language usage can make the task complex. #given_prompt#: <prompt>
Negation	Given a prompt #given_prompt# that represents two sentences and asks whether the two are paraphrases or not, based upon the #given_prompt# create a #new_prompt# in the same language by introducing negations or double negatives, the meaning of the sentence could be the same but the formation different. #given_prompt#: <prompt>
Time Navigation	Given a prompt #given_prompt# that represents two sentences and asks whether the two are paraphrases or not, based upon the #given_prompt# create a #new_prompt# in the same language by changing the time description (from past to present or future) in paraphrased sentences. #given_prompt#: <prompt>
Cultural Inferences	Given a prompt #given_prompt# that represents two sentences and asks whether the two are paraphrases or not, based upon the #given_prompt# create a #new_prompt# in the same language by using different cultural inferences in each sentence. The task gets complicated when two sentences infer same conclusion but uses culturally different examples or metaphors. #given_prompt#: <prompt>
Length Variation	Given a prompt #given_prompt# that represents two sentences and asks whether the two are paraphrases or not, based upon the #given_prompt# create a #new_prompt# in the same language by using different sentence length one can be short and another very long. #given_prompt#: <prompt>

Table 31: Pawsx

Evol Type	GPT-4 Prompt
Ambiguity	Given a prompt #given_prompt# based upon the #given_prompt# create a #new_prompt# in the same language but making it much more vague and ambiguous thus making it not so straightforward to answer. #given_prompt#: <prompt>
Long Form Question	Given a prompt #given_prompt# based upon the #given_prompt# create a #new_prompt# in the same language by making it longer i.e. formulating the questions in long and complex sentences thus requiring the system to decipher the main question. #given_prompt#: <prompt>
Multilingual	Given a prompt #given_prompt# based upon the #given_prompt# create a #new_prompt# in <language_2> have different linguistic structure. Finally, ask to answer the question in the <language_1>. #given_prompt#: <prompt>
Combine Facts	Given a prompt #given_prompt# based upon the #given_prompt# create a #new_prompt# in the same language by combining multiple facts thus making the question more complex and requiring combining multiple facts to answer correctly. #given_prompt#: <prompt>
Implicit Question	Given a prompt #given_prompt# based upon the #given_prompt# create a #new_prompt# in the same language by asking implicit question where the answer isn't explicit and requires understanding of the underlying implication. #given_prompt#: <prompt>
Negative Questions	Given a prompt #given_prompt# based upon the #given_prompt# create a #new_prompt# in the same language by asking negative questions that require the recognition of the negation included in the sentences. #given_prompt#: <prompt>
Inference Deduction	Given a prompt #given_prompt# based upon the #given_prompt# create a #new_prompt# in the same language by design question that require a degree of inference or deduction that might not be directly provided anywhere. #given_prompt#: <prompt>
Multiple Answers	Given a prompt #given_prompt# based upon the #given_prompt# create a #new_prompt# in the same language by design multiple-choice questions where more than one answer could be correct, making it more complex to find the right named entities. #given_prompt#: <prompt>
Comparitive Questions	Given a prompt #given_prompt# based upon the #given_prompt# create a #new_prompt# in the same language by asking questions would require the system to understand the comparative degree being asked about, making extraction or sorting from data more complicated. #given_prompt#: <prompt>

Table 32: Openqa

9.4 Prompt for Multiturn Evol-instruct

Evol Type	**GPT-4 Prompt**
Challenging	- The follow-up instruction must be challenging in terms of difficulty in comparison with the initial instruction.
Ambiguous	- The follow-up instruction must refer to the previous result obtained from the initial instruction in an ambiguous way (e.g., summarize that under 3 paragraphs...)
Redirection	- The follow-up instruction must abruptly change the type of the request/task or the thematic/topic of the initial instruction with no transition formula (e.g., let's shift gears) or even referring to the initial instruction.
Generic Rewriting	- The follow-up instruction must request a change in the {property} of the response to the INITIAL INSTRUCTION.
Feedback Handling	- The follow-up instruction must indicate that what the AI model responded to the INITIAL INSTRUCTION was not good enough (you must specify on which random aspect).
Random	- The follow-up instruction must request to change the response content or format in unique and unusual ways (e.g. switch to JSON or YAML or even a custom format illustrated by a template or very specific format description, keep all words starting with certain letter, remove every other word... You must specify this way in the instruction).
Context Retention	- The follow-up instruction must present a request/task that will test the ability of the model to retain the context of the conversation established by the previous instructions.
Format Rewriting	- The follow-up instruction must request a change in the format of the response to the previous instruction.
Persona Rewriting	- The follow-up instruction must request a change in the persona of the response to the previous instruction.
Detailed Constraints	- The follow-up instruction must add detailed constraints, like specifying the desired output format. Also involves providing more specific parameters or criteria to narrow down search results. Examples include specifying keywords, time ranges, locations, categories, or sources.
Adjust Output Format	- The follow-up instruction must ask to adjust the output format as users may request specific formats for the output, such as text-only, summarized results, or structured data formats.
Expanding Queries	- The follow-up instruction must ask to expand on a certain topic as users might want to broaden the search scope to include related topics or synonyms.
Refocus Queries	- The follow-up instruction must be a refocus query as users may wish to refocus the query to target a specific aspect or angle of their original request.
Change Context	- The follow-up instruction must introduce a new topic or context that is related to the current conversation, allowing the chatbot to provide a different perspective or information.
Clarification	- The follow-up instruction must ask for clarification as the chatbot may provide a complex or unclear response, ask for clarification to encourage it to expand on its answer.
Chatbot Opinion	- The follow-up instruction must encourage the chatbot to provide its own perspective or opinion on a topic, which can help create a more dynamic and engaging conversation.
Open Ended Questions	- The follow-up instruction must ask open-ended questions that require more detailed and thoughtful responses, encouraging the chatbot to provide more information and keep the conversation going.
Complex Queries	- The follow-up instruction must ask to create a multi-part question or instruction and see how the chatbot manages to break down and answer each part.
Pronouns	- The follow-up instruction must ask a question that uses pronouns like "it," "he," or "she" after some gap in the conversation. The bot should have to remember the noun the pronoun is referring to.
Engaging Conversation	- The follow-up instruction must engage the chatbot in a conversation about a topic that requires knowledge of previous interactions.
Recall Information	- The follow-up instruction must ask the chatbot to recall the details of the earlier turns in the conversation.

Table 33: Multiturn Evols

GPT-4 Multiturn Prompt

Your goal is to create a follow-up instruction to an INITIAL INSTRUCTION given to an AI model. You must design the follow-up using these specifications:

- The follow-up instruction must read like it's addressed to an AI model and not to another human. As such it should exclude requests impossible for an AI model to do (e.g. watch a movie or build a house).
- The follow-up instruction should be fully relevant and make sense regardless of the AI model's previous answer to the INITIAL INSTRUCTION. As such, it should rely on the INITIAL INSTRUCTION only and not on a hypothetical, unknown response by the AI model.
- The follow-up instruction should be in < *{language}* > and should be a natural continuation of the INITIAL INSTRUCTION.

{follow_up_type}

INITIAL INSTRUCTION: "*{instruction}* "

Provide directly the follow-up instruction requested with no additional comment, text or explanation, strictly in a valid json object:

{
"follow_up_user_prompt": "..."
}

Figure 4: Multiturn Prompt to GPT-4

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the claims made in the abstract and introduction are discussed throughout the paper and empirically shown in section 5 (Results).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we briefly describe limitations of our work in section 8.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We do not present theoretical assumptions or proofs in our paper, which is empirical in focus.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes] ,

Justification: We share the full taxonomy in appendix of this paper, describe steps in details (section 3) to generate **M2Lingual**, and provide all model hyperparameters (section 4.2).

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide all details for each aspect of the taxonomy based evols in appendix of this paper. We will share a small sample of the dataset and our code repository in supplementary material. We will share the full dataset and trained models with the camera-ready version of the paper.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe hyperparameters in section 4.2, dataset statistics in table 1, table 2, and captions or in-rows of each of the result tables.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We could not include the standard deviations in result tables in main content due to space constraints.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Mentioned in section 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We adhere to the NeurIPS Code of Ethics.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the ethical considerations along with limitations in Section 8.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We discuss the low risk of toxic or offensive data in our dataset in Section 8.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all relevant work that we utilize, and follow the respective licenses.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide documentation for reproducibility and our repository (provided in supplementary material) will also have documentation.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our data is fully synthetic and we did not employ human subjects to construct the datasets.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Since the dataset us fully synthetic and we do not conduct research with human subjects.