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

Rishabh Maheshwary[§], **Vikas Yadav**[§], **Hoang Nguyen**[†*]
Khyati Mahajan[§], **Sathwik Tejaswi Madhusudhan**[§]
[§] ServiceNow
[†] University of Illinois at Chicago
`{rishabh.maheshwary, vikas.yadav}` @servicenow.com[§]

Abstract

Collecting instruction fine-tuning (IFT) data is a resource and time intensive task, especially in multilingual settings where finding proficient native speakers is challenging. Moreover, traditional data collection is prone to privacy risks, toxicity and lacks scalability. While fully synthetic datasets are a promising alternative, research on their use in multilingual domain is limited as existing approaches still rely on machine translation to improve multilingual performance. To bridge this gap we introduce ***M2Lingual***, the first *fully synthetic*, *multi-turn* multilingual dataset having $175K$ conversations across 70 languages with a balanced mix of high, low and mid-resourced languages. ***M2Lingual*** is constructed using a cost-efficient and scalable method that uses our *novel two-step Evol* prompt taxonomy to transform a small set of human written instructions to complex and challenging conversations. Results across *three* model families, *six* baseline datasets and evaluation spanning 31 languages demonstrates the effectiveness of ***M2Lingual*** over other datasets.

1 Introduction

The recent success of large language models (LLMs) (Achiam et al., 2023; Jiang et al., 2024; Touvron et al., 2023a; Team et al., 2023) can be largely attributed to the availability of large, diverse, and high quality instruction fine-tuning (IFT) datasets (Taori et al., 2023; Chiang et al., 2023; Xu et al., 2023). However, the majority of IFT datasets are in English with very limited coverage for other languages (Zhang et al., 2023).

Existing multilingual IFT datasets can be divided into those that require human involvement and those that rely on machine translation (Table 1). The development of human-involved datasets is resource-heavy, often requiring native speakers, which introduces potential for annotator errors, uneven data distribution, and privacy and toxicity concerns (Abay et al., 2019; Zhao et al., 2024). These challenges lead to low-complexity conversations (Xu et al., 2023) as well. Machine-translated datasets offer less resource-intensive methods to create the data, but suffer from translation artifacts known as *translationese* (Bizzoni et al., 2020; Vanmassenhove et al., 2021) that fail to capture linguistic nuances (Wang et al., 2023a). In conjunction with limited language coverage, overly simple instructions, and unbalanced NLP task representation, most multilingual datasets are not multi-turn, limiting the ability of models to engage beyond single utterances (Wei et al., 2023).

Fully synthetic datasets offer a promising solution to address the above concerns. Not only do synthetic datasets address the high cost of data collection, toxicity and privacy concerns, english synthetic datasets like WizardLM, Vicuna, Ultrachat, etc have been proven to significantly enhance the performance of LLMs in English (Xu et al., 2023; Chiang et al., 2023; Ding et al., 2023). However, there is a lack of research on synthetic datasets in the multilingual domain that encompass a wide range of languages, NLP tasks, and multi-turn conversations. To address this gap, we present the following contributions:

1. We introduce ***M2Lingual***, the first *fully synthetic*, *multi-turn*, and *diverse* multilingual dataset, containing $175K$ complex and challenging conversations across $70+$ languages and 19 NLP tasks built with the *Evol* taxonomy.
2. We construct a *novel, two-step Evol* taxonomy (Figure 2), covering 19 NLP tasks, each with 9 distinct methods to transform seed instructions to make them more complex and challenging. Additionally, to synthesize multi-turn conversations, we develop 21 *Evol* prompts to increase engagement. This controlled setup *ensures a balances representation of different languages*,

---

[*] Work done during internship at ServiceNow

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 9676–9713
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

Dataset	Size	Multi turn?	Langs	Resource Level		Task specific?	General instructions?	Translated dataset?	Fully synthetic?
				Low	High				
Aya Dataset (Singh et al., 2024)	200K IR pairs	✗	70	37 (1)	32	✗	✓	✗	✗
MultiAlpaca (Wei et al., 2023)	132K IR pairs	✗	11	0	11	✗	✓	✗	✓
M-Alpaca (Chen et al., 2024)	52K IR pairs	✗	12	0	12	✗	✗	✓	✓
Bactrian-X (Li et al., 2023)	3.4M IR pairs	✗	52	15(1)	36	✗	✓	✓	✓
OpenAssistant (Köpf et al., 2024)	10K convs	✓	35	3	32	✗	✗	✗	✗
ShareGPT (RyokoAI, 2023)	94K convs	✓	45	4 (2)	39	✓	✗	✗	✗
WildChat (Zhao et al., 2024)	1.04M convs	✓	74	21 (3)	50	✗	✗	✗	✗
M2Lingual	182K convs	✓	70	37 (1)	32	✓	✓	✗	✓

Table 1: Comparison of multilingual IFT datasets with ***M2Lingual***. The top 4 rows are task based multilingual focused IFT datasets and the bottom 3 rows are datasets collected in the wild. Resource level classification taken from NLLB (Costa-jussà et al., 2022). Languages not found in the NLLB table are counted as low, in parentheses.

especially low resource languages (Figure 4) which is challenging to achieve in real-world scenarios (Przystupa and Abdul-Mageed, 2019). The *Evol* taxonomy enables a fully-synthetic, scalable, and cost-efficient method for constructing enriched multi-turn multilingual conversational IFT dataset which is extendable to any task and language.

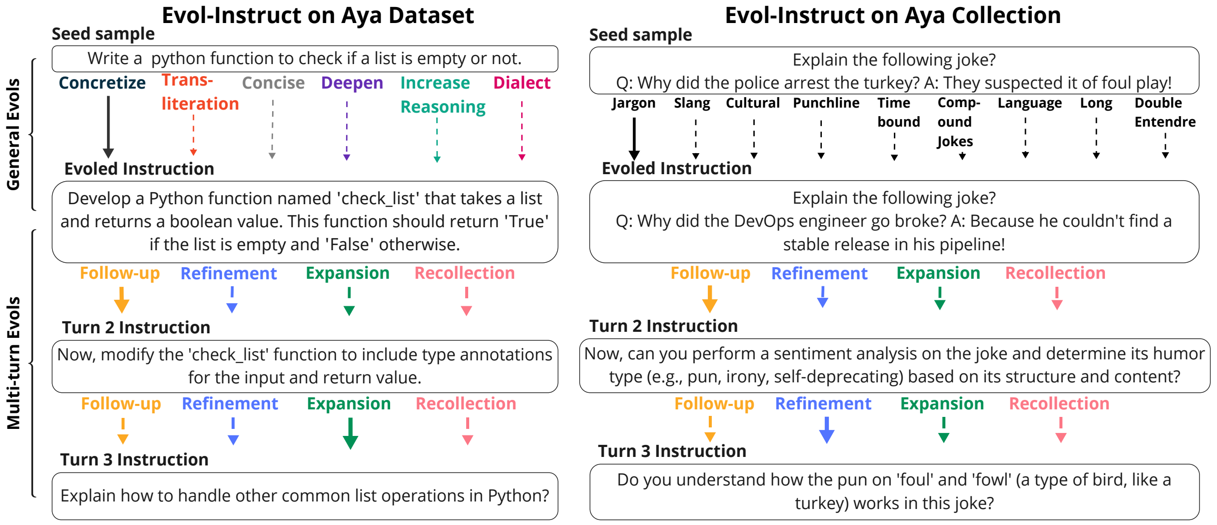
3. We provide detailed analyses highlighting the impact of seed instructions, each step of the data enrichment and synthesis process. Additional analysis on low resource languages, content moderation, conversation length, and language distribution, demonstrate the superiority of ***M2Lingual*** over other datasets.

2 Related Work

Multilingual Instruction Finetuning. Due to the widespread availability of high-resource language pretraining corpora multilingual instruction finetuning has proven to be a cost effective solution for improving performance (Ranaldi et al., 2023; Chen et al., 2024; Üstün et al., 2024). Several approaches have been adopted to expand access to multilingual IFT corpora. Notable among these are datasets derived from NLP tasks (e.g., FlanT5, Supernatural Instructions) (Chung et al., 2024; Sanh et al., 2021; Wang et al., 2022)

Human-generated datasets such as Aya (Singh et al., 2024) and OpenAssistant (Köpf et al., 2024) involve humans creating conversation topics, writing questions, and crafting responses. While these datasets are typically high quality, their creation is extremely resource and time intensive. Moreover, finding native speakers for diverse languages is challenging, with potential annotator errors and uneven data distribution. (Singh et al., 2024; Gilardi et al., 2023), making it difficult to scale these datasets. *Human-AI generated datasets* such as LM-Sys (Zheng et al., 2023a), WildChat (Zhao et al., 2024) and ShareGPT (RyokoAI, 2023) are less resource-intensive than purely human-generated ones, as they involve humans interacting with LLMs to generate conversations. However, they still present challenges, as humans must write instructions and create diverse questions in native languages, a process that remains time-consuming. Additionally, this approach can raise privacy concerns (Abay et al., 2019), introduce toxic data (Zhao et al., 2024), and result in low-complexity conversations (Xu et al., 2023). Finally, *machine-translated datasets* such as BactrainX (Li et al., 2023) offer a more resource efficient method. However, such datasets often suffer from translation artifacts known as *translationese* (Bizzoni et al., 2020; Vanmassenhove et al., 2021) that fail to capture linguistic nuances (Wang et al., 2023a). On the contrary to these, our presented ***M2Lingual*** dataset utilizes IFT seeds from native speakers across various languages (section 3.1) and applies task specific mutation in each language (section 3.2), thus maintaining linguistic nuances in respective individual language. ***M2Lingual***'s generation pipeline is also completely synthetic (table 1), making it a scalable and an affordable method for multilingual IFT data generation.

Synthetic Datasets. Fully synthetic datasets have emerged as a promising alternative towards addressing constraints with existing data generation methods. Popular English synthetic datasets, such as Alpaca (Taori et al., 2023), WizardLM (Xu et al., 2023), and Vicuna (Chiang et al., 2023), generate new instructions from a small initial set using methods like Self-instruct (Wang et al., 2023b) or *Evol*-Instruct (Xu et al., 2023), and have shown strong performance. However, there is limited research on leveraging synthetic datasets to enhance multi-



Write a python function to check if a list is empty or not.

Develop a Python function named 'check_list' that takes a list and returns a boolean value. This function should return 'True' if the list is empty and 'False' otherwise.

Now, modify the 'check_list' function to include type annotations for the input and return value.

Explain how to handle other common list operations in Python?

Explain the following joke?
Q: Why did the police arrest the turkey? A: They suspected it of foul play!

Explain the following joke?
Q: Why did the DevOps engineer go broke? A: Because he couldn't find a stable release in his pipeline!

Now, can you perform a sentiment analysis on the joke and determine its humor type (e.g., pun, irony, self-deprecating) based on its structure and content?

Do you understand how the pun on 'foul' and 'fowl' (a type of bird, like a turkey) works in this joke?

Figure 1: Walk-through for data synthesis of ***M2Lingual***. Step 1 is seed selection. In Step 2 for each instruction corresponding task specific *Evol* prompt taxonomy is used for generating complex evoled instruction. Finally, in Step 3, multi-turn instruction are generated on Step 2 evoled instructions using multi-turn *Evol* prompt taxonomy.

lingual capabilities, with the exception of MultiAlpaca (Wei et al., 2023), which uses Self-instruct. This approach has been shown to be susceptible to repetitive and noisy outputs (Chen et al., 2023; Ghosh et al., 2024), and suffers from low performance (Tables 12 & 13).

3 Methodology

M2Lingual has three main synthesis steps. *Step 1: Seed Selection* involves the selection of diverse multilingual seeds. *Step 2: Guided Evol* uses the *Evol* prompt taxonomy to generate complex instruction and response (IR) pairs and *Step 3: Multi-turn Evol* uses the multi-turn portion of the taxonomy to extend IR pairs to multilingual conversations. Figure 1 captures an overview of each step in ***M2Lingual*** synthesis and Figure 2 presents the categories of *Evol* prompts[*].

3.1 Seed Selection

To ensure that we select diverse seeds capturing language nuances and covering a variety of NLP tasks, we select seed examples from the Aya dataset and collection as both have a high average approval ratio by human annotators (Singh et al., 2024).

Aya dataset seeds. Aya dataset has general IR pairs written by native speakers that captures region specific language nuances and cultural contexts. We randomly select 100 IR pairs for each of the 70 languages, resulting in 7000 seed IR pairs.

Aya collection seeds. Aya collection covers 19 NLP tasks where each task has parallel examples in 113 languages. To ensure a proper balance of the number of examples across all languages, we only focus on 70 languages and exclude two NLP tasks — 1) text simplification, as it is already supported by our *Evol* prompts, and 2) multilingual event entity task, as Aya do not have a consistent format for this task. Finally, for each task in the collection, we randomly sample 6 examples per language, resulting in $6 \times 70 \times 17 = 7140$ IR seeds. 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 seeds contain $7000+7140 = 14140$ IR samples.

3.2 Guided *Evol*

The seed instructions span a variety of NLP tasks but are generally straightforward and overly simplistic. To enhance LLMs' instruction following capabilities, particularly for complex tasks, we apply *Evol-Instruct* (Xu et al., 2023) to our selected seed instructions. The *Evol-Instruct* method uses Evol prompts to transform simple instructions into more intricate ones. However, the generic Evol conditions from the original work[†] provide very limited guidance for generating new IR pairs, especially for the 19 diverse NLP tasks for which we aim to generate training data. To address this, we

---

[*] Complete *Evol* taxonomy prompts are in Appendices A.10 and A.11

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

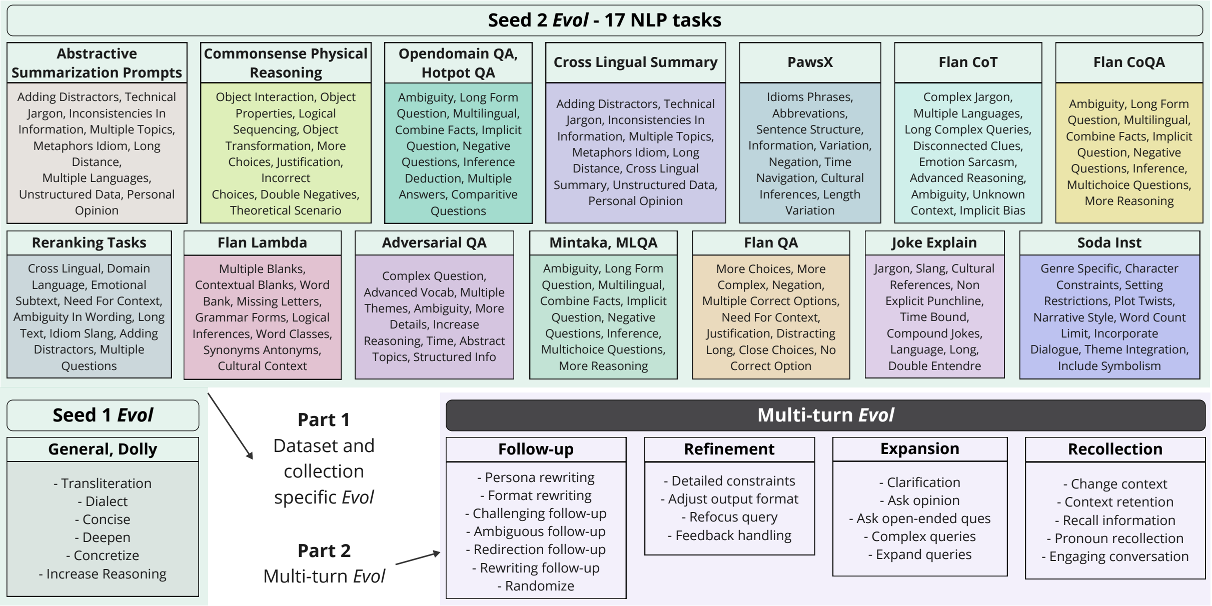

Seed 2 Evol - 17 NLP tasks
Abstractive Summarization Prompts
Adding Distractors, Technical Jargon, Inconsistencies In Information, Multiple Topics, Metaphors Idiom, Long Distance, Multiple Languages, Unstructured Data, Personal Opinion
Commonsense Physical Reasoning
Object Interaction, Object Properties, Logical Sequencing, Object Transformation, More Choices, Justification, Incorrect Choices, Double Negatives, Theoretical Scenario
Opendomain QA, Hotpot QA
Ambiguity, Long Form Question, Multilingual, Combine Facts, Implicit Question, Negative Questions, Inference Deduction, Multiple Answers, Comparitive Questions
Cross Lingual Summary
Adding Distractors, Technical Jargon, Inconsistencies In Information, Multiple Topics, Metaphors Idiom, Long Distance, Cross Lingual Summary, Unstructured Data, Personal Opinion
PawsX
Idioms Phrases, Abbrevations, Sentence Structure, Information, Variation, Negation, Time Navigation, Cultural Inferences, Length Variation
Flan CoT
Complex Jargon, Multiple Languages, Long Complex Queries, Disconnected Clues, Emotion Sarcasm, Advanced Reasoning, Ambiguity, Unknown Context, Implicit Bias
Flan CoQA
Ambiguity, Long Form Question, Multilingual, Combine Facts, Implicit Question, Negative Questions, Inference, Multichoice Questions, More Reasoning
Reranking Tasks
Cross Lingual, Domain Language, Emotional Subtext, Need For Context, Ambiguity In Wording, Long Text, Idiom Slang, Adding Distractors, Multiple Questions
Flan Lambda
Multiple Blanks, Contextual Blanks, Word Bank, Missing Letters, Grammar Forms, Logical Inferences, Word Classes, Synonyms Antonyms, Cultural Context
Adversarial QA
Complex Question, Advanced Vocab, Multiple Themes, Ambiguity, More Details, Increase Reasoning, Time, Abstract Topics, Structured Info
Mintaka, MLQA
Ambiguity, Long Form Question, Multilingual, Combine Facts, Implicit Question, Negative Questions, Inference, Multichoice Questions, More Reasoning
Flan QA
More Choices, More Complex, Negation, Multiple Correct Options, Need For Context, Justification, Distracting Long, Close Choices, No Correct Option
Joke Explain
Jargon, Slang, Cultural References, Non Explicit Punchline, Time Bound, Compound Jokes, Language, Long, Double Entendre
Soda Inst
Genre Specific, Character Constraints, Setting Restrictions, Plot Twists, Narrative Style, Word Count Limit, Incorporate Dialogue, Theme Integration, Include Symbolism
Seed 1 Evol
General, Dolly
- Transliteration
- Dialect
- Concise
- Deepen
- Concretize
- Increase Reasoning
Part 1
Dataset and collection specific Evol
Part 2
Multi-turn Evol
Multi-turn Evol
Follow-up
- Persona rewriting
- Format rewriting
- Challenging follow-up
- Ambiguous follow-up
- Redirection follow-up
- Rewriting follow-up
- Randomize
Refinement
- Detailed constraints
- Adjust output format
- Refocus query
- Feedback handling
Expansion
- Clarification
- Ask opinion
- Ask open-ended ques
- Complex queries
- Expand queries
Recollection
- Change context
- Context retention
- Recall information
- Pronoun recollection
- Engaging conversation

Figure 2: Taxonomy of *Evol* prompt conditions applied towards creating ***M2Lingual***. Part 1 includes *Evol* prompts for Aya seeds and Part 2 has multi-turn *Evol* prompts applied for creating conversation.

develop a novel *Evol* prompt taxonomy covering both general instructions and specific NLP tasks, as illustrated in Figure 2.

***Evol* Prompt Taxonomy.** For general instructions from the Aya dataset, we design 6 *Evol* prompt conditions that enhance multilingual features. We create 9 task-specific *Evol* prompts for each NLP task in the Aya collection to ensure that we tailor *Evol* conditions for individual tasks. We use GPT-4 (Achiam et al., 2023) to transform the seeds using our *Evol* prompt taxonomy. These are applied to specific seeds as described below:

- *Aya dataset seeds.* As Aya dataset has general IR pairs, we apply the 6 *generic evol prompts* to each seed example. This results in $7K \times 6 = 42K$ instructions which are complex, challenging, and captures all nuances and complexities of languages.
- *Aya collection seeds.* For each seed instruction from one of the 17 tasks (top block of Figure 2), we apply its corresponding 9 *Evol*s resulting in a total of $7140 \times 9 = 64260$ instructions from Aya collection.

3.3 Multi-turn *Evol*

The final step involves generating multiple user-assistant turns based on the task- *Evol*ed instructions from the previous phase. Conversations between a user and an AI assistant generally fall into four broad categories: *Follow-up*, *Refinement*, *Expansion*, and *Recollection* (Kwan et al., 2024). We propose a multi-turn *Evol* prompt taxonomy with 21 distinct dialogue variations (final block labeled as part 2 in Figure 2) that build upon the original generic four categories. Additionally, we ensure that all subsequent instructions are generated in the same language as the initial instruction by explicitly prompting GPT-4. We select all the *Evol*ed instructions from the Aya dataset, and pick a balanced subset of size 35K from Aya collection and generate turns as follows:

1. *User turns.* We use the prompt specified in Appendix A.11 to generate multiple user turns. Specifically, we use the task-*Evol*ed instruction from the previous step (i.e., Step 2), with its language and one of the 21 dialogue variations to generate the next user instruction.
2. *Assistant turns.* For all the generated user turns, we generate subsequent responses via GPT-4 using the entire conversation history. To mitigate the impact of topic drift from the long conversations (Zhang and Zhao, 2021), we restrict the total number of turns to $<= 6$.

Post-Hoc Filtering. Upon manual inspection, we find that some IR pairs generated using GPT-4 have repetitive long sequences and n-grams. To mitigate this, we apply a filtering step following (Guo et al., 2021; Elmadany et al., 2023) to remove IR pairs with frequent n-grams. This filtering is performed after steps 2 and 3. The final dataset consists of 75K multi-turn conversations with 100K single turn conversations.

4 Experiments

We conduct experiments across *three* model families & *five* model sizes — Mistral-7B (Jiang et al., 2023), LLaMA-3-8B (Touvron et al., 2023a) and QWEN-4B (Bai et al., 2023). Furthermore, to demonstrate the effectiveness across different model scales, we fine-tune a larger model, LLaMA-2-13B (Touvron et al., 2023b), and a smaller model, QWEN-1.8B (Bai et al., 2023). To evaluate how well the datasets work with instruction-tuned models, we also experiment with Mistral-Instruct-7B.

Baselines — We use 3 multilingual instruction fine-tuning (IFT) datasets *MultiAlpaca*, *Bactrian-X*, and *Aya* for main evaluation. Furthermore, to highlight the importance of each step in our synthesis, we consider several ablations. Specifically, we train models using 1) only **Seed** samples, 2) seed samples with the generated *Evol* s (**Seed + Evol**) and 3) seeds, *Evol* s and the multi-turn conversations (**Seed + Evol + MT**).

4.1 Evaluation

Multilingual benchmarks. We utilize the EleutherAI evaluation (Gao et al., 2021) for consistent comparisons on the following tasks:

- *Question Answering*: We focus on 3 QA datasets 1) XQUAD (Artetxe et al., 2019), TyDiQA (Clark et al., 2020) and MLQA (Lewis et al., 2019). We use 3 in-context examples and 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 with F1-score as the metric.
- *Summarization*: We use the XLSUM (Hasan et al., 2021) dataset and focus on 6 languages — Arabic, English, Spanish, French, Japanese and Russian with 100 examples per language and use GPT-4 as a judge to rate the generated summaries on a scale of 1 to 5. We also report ROUGE$_L$ (Lin, 2004) & BLEU (Papineni et al., 2002) for comparison.
- *Multilingual math word problems*: We use MGSM (Shi et al., 2022) that translates GSM8K (Cobbe et al., 2021) to 10 languages. We use 3 in-context examples and compute exact match (EM) with ground truth answer.
- *Classification*: We focus on XNLI (Conneau et al., 2018) and XCOPA (Ponti et al., 2020) with 15 and 11 languages respectively in a zero-shot setting and report the resuls in Appendix. We compute the accuracy (Acc) by looking at the log-likelihood assigned to the ground truth answer on the validation set.

Multilingual MT-Bench. We evaluate conversational complex instruction following ability using MT-Bench (Zheng et al., 2023b). It has 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 responses on a scale of 1 to 10, with the overall score being the mean over two turns. We translate it into 8 different languages (French, Canadian French, German, Italian, Spanish Japanese, Dutch, Portuguese) with professional linguists to ensure high quality evaluation. We modify the judge prompt to include the language of the question, and instruct GPT-4 to make sure the responses are in the same language. We report the average scores across 80 examples for each language and the average MT-Bench score across all languages.

Low-resource Evaluation. To demonstrate the wide coverage of low resource languages in ***M2Lingual***, we further evaluate models by on by translating MT-Bench to 6 low-resource languages namely Hindi, Urdu, Thai, Tamil, Bengali and Gujarati using GPT-4. Finally, we also perform low-resource evaluation across 10 languages from Flores200 (Costa-jussà et al., 2022; Goyal et al., 2022). We present BLEU (Papineni et al., 2002) scores for translating each language into every other language. The final score for a language is the average BLEU score across all its translations to the remaining languages. The languages we used are Arabic (arb), Assamese (asm), Awadhi (awa), Belarusian (bel), Haitian Creole (hat), Kirghiz (kir), Burmese (mya), Nepali (nep), Somali (som), and Yoruba (yor). This selection covers a wide range of geographic regions (South Asia, Africa, Eastern Europe, and the Middle East) and includes languages with different writing systems: Latin, Cyrillic, Arabic, and Devanagari scripts.

5 Results

Multilingual MT-Bench — Table 2 shows the average scores on MT-Bench across 9 languages. On average, ***M2Lingual*** outperforms other baseline datasets by 1.01 MT-bench score with Mistral-7B, and 1.2 with LLama-3-8B. The significant improvements across most baselines highlight the

Model	Dataset	MT-AVG	XQUAD F1	TyDiQA F1	MLQA F1	XLSUM GPT-4 score	MGSM EM
Mistral-7B	MultiAlpaca	4.77	67.99	64.44	55.69	3.06	11.71
	Bactrian-X	5.25	71.91	66.63	60.27	2.59	17.14
	Aya	5.18	70.46	66.95	57.47	3.13	13.86
	Seed	5.04	72.52	65.89	59.33	2.94	13.65
	Seed + Evol	5.56	71.01	65.04	57.47	3.41	18.38
	Seed + Evol + MT (***M2Lingual***)	**6.54**	**74.53**	**67.57**	**62.40**	**3.55**	**22.00**
LLaMA-3-8B	MultiAlpaca	4.41	75.08	64.49	59.01	3.17	10.68
	Bactrian-X	5.66	69.57	56.45	58.51	2.79	22.86
	Aya	4.95	75.14	59.60	53.14	3.29	22.09
	Seed	3.54	**77.27**	68.57	60.01	2.84	11.71
	Seed + Evol	6.12	76.17	**69.89**	63.09	3.52	**28.00**
	Seed + Evol + MT (***M2Lingual***)	**6.74**	75.91	67.84	**63.5**	**3.68**	27.36

Table 2: Performance comparison of Mistral-7B and LLaMA-3-8B. MT-Avg is average MT bench results across 9 languages (French, Canadian French, German, Italian, Spanish, Japanese, Dutch, Portuguese). *Seeds* are 15.1K seeds; *Seed + Evol* is additional *Evol* IR pairs. *Seed + Evol + MT* has additional multi-turn data.

strengths of ***M2Lingual***. Detailed results across all 9 languages of MT-bench with different base models are shown in Appendix in Table 12 and 15.

Multilingual NLP results — ***M2Lingual*** leads in performance on 3 multilingual NLP tasks across Mistral-7B and LLama-3-8B. Specifically in Table 2, on average ***M2Lingual*** always outperforms on QA task by $6-8\%$ and on MGSM by 8% across both models. On summarization across both models our dataset outperforms by 0.5 GPT-4 score on an average. We observed GPT-4 score to be a more reliable metric than BleU or ROUGEL score for evaluating summarization quality, as those metrics tend to un-fairly penalize long form LLM answers (more in Section 6). However for completeness we report results on BleU and ROUGEL in Table 13. Finally, Table 13 and 14 in Appendix has results on XNLI, XCOPA and other base models Appendix A.8.

Importance of *Evol* — The **Seed + *Evol*** rows in Table 2 show the performance of seeds and synthetic *Evol* IR pairs. In comparison to **Seed** data across both models, the average MT-Bench score increases by at least 1.75 points, with gains in every language (see Table 12 and 15), especially Japanese. Similarly, ***M2Lingual*** leads to improvements of around 5 and 10 points on math word problem across Mistral-7B and LLama-3-8B respectively. On summarization and QA, the results are mixed (either increase, slight drop or same) as shown in Table 2. This is due to the increased verbosity of LLMs trained on ***M2Lingual***, which contains detailed synthetic *Evol*s. (more in Section 6). Finally, to demonstrate that the importance lies in the use of synthetic *Evol*s rather than simply increasing the amount of seed-like data, we sampled an additional $86K$ IR pairs from the Aya dataset and collection and replace it with synthetic generated *Evol*s and observed that the performance decreases across all benchmarks, especially in MT-Bench and MGSM by 1.1 and 3.38 points respectively (results are shown in Appendix A.5).

Importance of multi-turn *Evol*. The **Seed + *Evol* + MT** rows in Table 2 shows performance after adding synthetically generated turns using multi-turn *Evol*. This boosts performance in MT-Bench evaluations substantially by 0.8 points with the most significant gain of 1.31 and 1.0 points on French and Japanese for Mistral-7B model (Table 12). Adding multi-turn data also helps in multilingual benchmarks as the results consistently improve by $3-4\%$ across all evaluations. Additional results are shown in Appendix, Table 12, 13, 14, and, 15.

6 Additional Analysis

Comparison with Human-AI generated data in the wild. For completeness, we also include performance comparisons with Human-AI generated datasets collected from voluntary participation, such as OpenAssistant, ShareGPT, and WildChat in Table 3, where ***M2Lingual*** shows strong performance results with both Mistral-7B and Llama-3-8B. Concretely, ***M2Lingual*** outperforms OpenAssistant and ShareGPT by 0.8 and 0.6 on multilingual MT-Bench, 8% and 12% on QA, 0.7 and 1.0 on summarization and 8% and 5%

Model	Dataset	MT-AVG	XQUAD F1	TyDiQA F1	MLQA F1	XLSUM GPT-4 score	MGSM EM
Mistral-7B	Open Assistant	5.66	67.99	54.22	53.64	2.60	16.05
	ShareGPT	5.80	66.33	56.97	50.78	2.41	11.32
	WildChat	6.53	72.55	64.27	59.53	2.93	18.41
	M2Lingual	**6.54**	**74.53**	**67.57**	**67.57**	**3.55**	**22.00**
LLaMA-3-8B	Open Assistant	5.12	64.38	52.65	47.08	2.92	17.36
	ShareGPT	6.10	56.98	58.48	43.43	2.54	25.32
	WildChat	**6.75**	63.15	59.88	63.16	2.90	26.36
	M2Lingual	6.74	**75.91**	**67.84**	**63.50**	**3.68**	**27.36**

Table 3: Performance comparison of Mistral-7B and LLaMA-3-8B with OpenAssitant, ShareGPT & WildChat.

Model	Dataset	MT_bn	MT_gu	MT_hi	MT_ur	MT_th	MT_ta
Mistral-7B	MultiAlpaca	1.54	1.28	2.38	1.60	2.73	1.38
	Bactrian-X	3.58	2.75	3.93	3.21	4.01	2.37
	Aya	2.46	1.42	2.44	2.27	2.15	2.01
	Wildchat	2.23	1.22	3.37	2.34	3.28	1.53
	M2Lingual	**3.92**	**3.3**	**4.52**	**3.80**	**4.01**	**2.57**
LLaMA-3-8B	MultiAlpaca	2.68	2.35	3.08	2.33	3.25	2.09
	Bactrian-X	3.13	2.75	4.13	2.63	3.85	2.13
	Aya	3.3	2.4	3.9	2.86	3.57	2.68
	Wildchat	4.52	3.6	5.44	4.27	5.30	4.16
	M2Lingual	**4.73**	**3.91**	**5.97**	**4.5**	**5.68**	**4.25**

Table 4: Low-resource evaluation of Aya, WildChat, and ***M2Lingual*** using Mistral-7B and LLama-3-8B base models on Bengali (bn), Gujarati (gu), Hindi (hi), Urdu (ur), Thai (th), and Tamil (ta).

on math word problem solving. ***M2Lingual*** performs comparable to WildChat on multilingual MT-Bench however strongly outperforms by $2-3\%$ on QA, 0.6 on summarization and 2% on math word problem solving. We report per language MT bench scores, results on other metrics and classification benchmarks in Appendix (Tables 12, 13 14 and 15). Finally we also compare the performance of ***M2Lingual*** on low resource languages (see below) and observed that ***M2Lingual*** notably outperforms these Human-AI generated datasets as shown in table 4 and table 5. However, we would like to point out that these datasets are not focused specifically towards improving multilingual abilities, even though their creation methods lend inadvertently to multilingual data. For holistic comparison, we also include details with the next best performing dataset WildChat in further analysis.

Low-resource languages. Table 4 shows the results on 6 languages. ***M2Lingual*** performs better than all the baselines across both Mistral-7B and LLama-3-8B. Specifically, ***M2Lingual*** improves the performance by 1.3 and 0.2 on average for both models respectively. We also evaluate cross-lingual machine translation performance on extremely low resource languages as shown in Table 5. Table 5 demonstrates that ***M2Lingual*** outperforms existing baseline datasets by noticeable margin of 0.2 BLeU score improvements with Mistral-7B and 0.5 with LLama-3-8B. The M2Lingual models outperform all baselines in translating between different low-resource languages, except on Somhali, Awadhi (LLama-3-8B) and Yoruba (Mistral-7B) where their performance is a close second. This highlights a better coverage of low-resourced languages in ***M2Lingual*** (Figure 4).

Effect of IFT datasets on different sized LLMs. We also study the impact of ***M2Lingual*** on a smaller scale model (QWEN-1.8B) and a larger model (LLaMA-2-13B). As shown in Table 6, QWEN-1.8B on an average, ***M2Lingual*** leads to 1.96, 10 and 4.5 points improvements across MT-bench, QA and MGSM respectively. Similarly for the LLaMA-2-13B we get 0.8, 3.75 and 3.0 points increase. The MT-Bench results across different languages are shown in Table 16 in Appendix.

Scaling data synthesis with seed size. Finally, to show how performance changes as we scale synthetic data generation on more seed examples only, we ran 2 ablations where we 1) use only 25% of seed examples and use its synthesized data and 2) use 50% of the seeds. Figure 3 shows that as

Model	Dataset	arb	asm	awa	bel	hat	kir	mya	nep	som	yor	Avg
Mistral-7B	Multialpaca	0.72	0.47	0.53	0.46	0.66	0.41	0.37	0.59	0.89	0.79	0.58
	Bactrian-X	0.57	0.6	0.96	0.68	0.6	0.52	0.39	0.78	0.63	0.45	0.61
	Aya	0.86	0.52	0.64	0.74	1.42	0.61	0.41	0.57	0.83	**1.38**	0.79
	WildChat	0.5	0.48	0.66	0.62	0.97	0.76	0.47	0.67	0.68	0.76	0.65
	M2Lingual	**1.31**	**0.94**	**0.91**	**0.82**	**1.57**	**0.83**	**0.56**	**1.2**	**1.05**	1.01	**1.02**
LLama-3-8B	Multialpaca	1.74	0.88	1.04	0.84	1.29	1.02	0.74	1.34	0.96	1.2	1.1
	Bactrian-X	1.51	0.89	1.02	1.13	1.23	1.00	0.41	1.09	1.13	0.88	1.02
	Aya	2.07	1.11	2.3	1.49	1.69	1.23	0.96	1.12	**1.66**	1.55	1.5
	WildChat	2.11	1.61	**2.27**	1.18	2.12	1.24	1.13	1.45	1.12	1.31	1.5
	M2Lingual	**2.9**	**1.93**	1.99	**2.15**	**3.27**	**1.76**	**1.35**	**2.15**	1.37	**1.77**	**2.06**

Table 5: Low-resource evaluation across 10 languages from Flores200. We present BLEU scores for translating each language into every other language. The final score for a language is calculated as the average BLEU score across all its translations to the remaining languages.

Model	Dataset	MT-AVG	XQUAD F1	TyDiQA F1	MLQA F1	XLSUM GPT-4 score	MGSM EM
Qwen-1.8B	MultiAlpaca	2.03	31.60	33.38	19.30	2.03	7.45
	WildChat	2.59	45.12	42.42	29.39	1.90	8.00
	M2Lingual	**4.27**	**52.12**	**47.66**	**38.24**	**2.83**	**12.23**
LLaMA-2-13B	MultiAlpaca	4.46	55.07	59.46	48.74	2.48	7.80
	WildChat	6.00	67.64	60.14	53.69	2.55	9.95
	M2Lingual	**6.08**	**69.38**	**64.66**	**54.64**	**3.03**	**11.95**

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

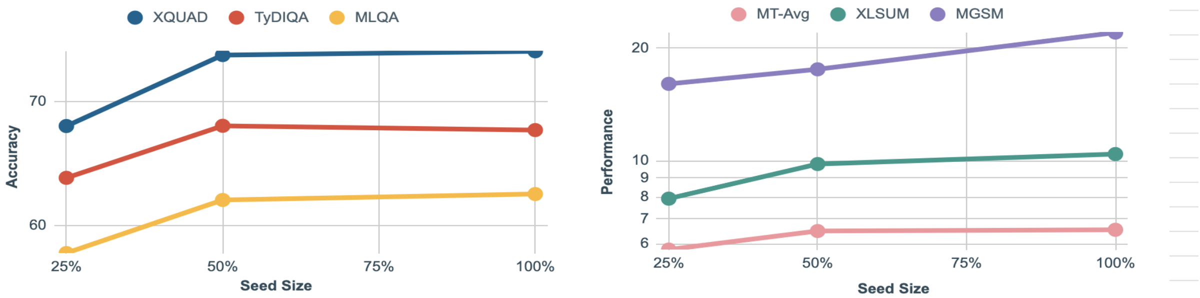


Figure 3: Performance vs seed size in data synthesis

we scale data synthesis by selecting more seeds the performance increases across all benchmarks. Specifically, on an average we see 0.5 improvement in multilingual MT-bench, 5% in QA and MGSM and 2.50 in summarization. Additional analysis and tables for Figure 3 are shown in Appendix A.6.

Distribution of Languages. Figure 4 shows a balanced representation of languages in our dataset compared to WildChat and Aya, which have uneven or very skewed distribution. This highlights a broader coverage of mid to low resource languages and explains the consistent performance improvements across high-mid resource languages in Tables 2, 3, 12 and 13 and low resource languages in Table 4 and 5.

Token Lengths per Utterance. Table 7 shows that ***M2Lingual*** has one of the highest user and assistant turn tokens and the highest total number of tokens (computed via LLama tokenizer). This explains on complex benchmarks such as MT-Bench and MGSM, which require reasoning using a chain-of-thought or processing long contexts; and an occasional slight drop in performance on QA or summarization tasks, as the F1-score does not fully capture long, detailed answers.

Content moderation. To ensure low toxicity in ***M2Lingual***'s content as well as evaluate the *Evol* synthesis method's sensitivity in data generation, we conduct moderation testing with OpenAI Moderation API (OpenAI, 2024) following (Zhao et al., 2024). Table 8 shows that less than 0.2%

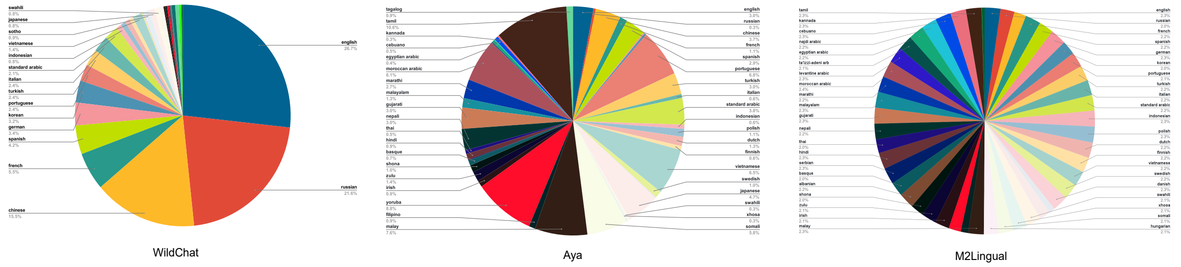


Figure 4: Comparison between Aya, WildChat and ***M2Lingual*** language distribution.

Dataset	#User tokens	#Assistant	# Total
Aya	157.9	447.2	605.1
Bactrian X	114.34	354.07	468.41
Multialpaca	47.58	97.22	144.8
Open Assistant	37.54	277.22	314.7
ShareGPT	99.97	558.06	658.03
WildChat	**282.33**	442.25	724.58
M2Lingual	200.28	**558.86**	**759.14**

Table 7: Token statistics for different datasets

of ***M2Lingual*** is flagged by the Moderation API. We remove the flagged utterances before making the dataset public. It is worth noting that Human-AI generated datasets like WildChat have substantially more sensitive content.

Dataset	User%	Chatbot%	Avg%
Alpaca	0.01	0.02	0.01
Aya	0.10	0.24	0.22
Open Assistant	0.53	0.45	0.49
Share GPT	0.16	0.28	0.22
WildChat	6.05	5.18	5.61
M2Lingual	0.21	0.13	0.17

Table 8: Content moderation analysis reported from respective dataset papers (BactrainX does not perform toxicity analysis)

7 Conclusion

We introduce ***M2Lingual*** - the *first fully synthetic, multi-turn multilingual dataset* - containing $175K$ complex conversations across 70+ languages and 19 NLP tasks. We propose a scalable, cost-efficient and fully synthetic method for creating conversations using a two-step enrichment process based on the *Evol* prompt taxonomy, which can be adapted to any task or monolingual data. Exhaustive experiments across *three* model families and *five* model sizes with evaluations spanning 31 languages demonstrate the advantages of ***M2Lingual*** over other datasets. Furthermore, our ablations and analysis on low-resource language support, content moderation, conversation length and language distribution demonstrate the quality of ***M2Lingual*** over other datasets.

8 Limitations and Ethical Considerations

M2Lingual covers over 70 languages in total, with dialects added in with *Evol* as well - a significant number of languages, more than all relevant datasets (Table 1). However, there are many more languages in the real-world, and we cannot cover them all. We hope that our contribution helps expand access to languages, and future work can further build better access for all. Moreover, the performance of LLMs improves on low resource data with finetuning on ***M2Lingual***, showcasing the importance of including multiple languages and turns.

Some major limitations of ***M2Lingual*** include the limited conversation length, possible presence of toxic data, and dependence on GPT-4 translated MT-Bench for low-resource language evaluation. While potentially longer conversations could be built with *Evol*, it would take significantly more resources to extend each conversation beyond the current limit. For toxicity, our seed dataset Aya does not contain specific flags for toxic, harmful, or offensive speech (Singh et al., 2024), and Aya authors report that they believe there is a low risk for these in Aya data. However, to mitigate risk, we conduct moderation analysis of the generated *Evol* IR pairs for ***M2Lingual***, and find that less than 0.2% of the generated data was flagged, which we filter out before making the data public. Lastly, we conduct limited manual evaluation of the GPT-4 generated low-resource multilignual MT-Bench data generated by GPT4, and find that it performs satisfactorily well. However, improving evaluation on low-resource data remains an area of future work.

References

Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. 2019. Privacy preserving synthetic data release using deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 510–526. Springer.

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Adeola Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International conference on spoken language translation*, pages 280–290.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa

Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *The 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Jack Choquette, Wishwesh Gandhi, Olivier Giroux, Nick Stam, and Ronny Krashinsky. 2021. Nvidia a100 tensor core gpu: Performance and innovation. *IEEE Micro*, 41(02):29–35.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm. *Company Blog of Databricks*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Octopus: A multitask model and toolkit for arabic natural language generation. *arXiv preprint arXiv:2310.16127*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Deepali Aneja, Zeyu Jin, Ramani Duraiswami, Dinesh Manocha, et al. 2024. A closer look at the limitations of instruction tuning. *arXiv preprint arXiv:2402.05119*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Han Guo, Bowen Tan, Zhengzhong Liu, Eric P Xing, and Zhiting Hu. 2021. Efficient (soft) q-learning for text generation with limited good data. *arXiv preprint arXiv:2106.07704*.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

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

OpenAI. 2024. Openai moderation api.

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

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Michael Przystupa and Muhammad Abdul-Mageed. 2019. Neural machine translation of low-resource and similar languages with backtranslation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 224–235.

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

RyokoAI. 2023. Sharegpt.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. 2021. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. *arXiv preprint arXiv:2102.00287*.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F Chen. 2023a. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *arXiv preprint arXiv:2309.04766*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. 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)*, pages 13484–13508.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Supernaturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of* *the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2023. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Zhuosheng Zhang and Hai Zhao. 2021. Advances in multi-turn dialogue comprehension: A survey. *arXiv preprint arXiv:2103.03125*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2023a. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Appendix

A.1 Experiment Details

We conduct experiments across *three* model families & *five* model sizes — Mistral-7B (Jiang et al., 2023), LLaMA-3-8B (Touvron et al., 2023a) and QWEN-4B (Bai et al., 2023). Furthermore, to demonstrate the effectiveness of our dataset across different model scales, we fine-tune both a larger model, LLaMA-2-13B (Touvron et al., 2023b), and a smaller model, QWEN-1.8B (Bai et al., 2023). To evaluate how well the datasets work with instruction-tuned models, we also experiment with Mistral-Instruct-7B.

A.2 Baseline Datasets

We use *six* different multilingual datasets as baselines for comparison: 1) the top ranked conversation trees from **Open Assistant** (Köpf et al., 2024), 2) **Aya** (Singh et al., 2024), 3) self-instruct dataset **MultiAlpaca** (Wei et al., 2023), 4) machine translated **Bactrian-X** (Li et al., 2023) derived from Alpaca-52k (Taori et al., 2023) and Dolly-15k (Conover et al., 2023), 5) the **ShareGPT** [‡] collection, and 6) **WildChat** (Zhao et al., 2024).

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 proportions as in 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**).

A.3 Training

All training is performed on 8 A-100 80GB NViDIA GPUs (Choquette et al., 2021), with the Axolotl[§] framework. We used Mistral tags (Jiang et al., 2023) for finetuning all models. We use a batch size of 64, a maximum sequence length of 8192, a learning rate of 5×10^{-6}, the Adam optimizer (Kingma and Ba, 2014) 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.

A.4 Evaluation

Multilingual benchmarks. We utilize the EleutherAI evaluation framework (Gao et al., 2021) 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 (Artetxe et al., 2019) with QA across 11 languages, 2) TyDiQA (Clark et al., 2020) which has human generated QA in 11 languages and 3) MLQA (Lewis et al., 2019) 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 (Hasan et al., 2021) 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 (Lin, 2004) & BLEU (Papineni et al., 2002) scores for comparison.
- *Classification*: We focus on XNLI (Conneau et al., 2018) and XCOPA (Ponti et al., 2020) 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 (Shi et al., 2022), a grade-school math benchmark that translates GSM8K (Cobbe et al., 2021) 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 multi-

---

[‡] https://sharegpt.com/

[§] https://github.com/OpenAccess-AI-Collective/axolotl lingual models across a wide array of tasks and languages, we translate MT-Bench (Zheng et al., 2023b). 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.

Model	Benchmark	**Aya-seeds**	**Seed + Evol**
Mistral-7B	MT-Avg	4.40	**5.57**
	XQUAD	70.40	**71.01**
	MLQA	56.10	**57.47**
	MGSM	15.32	**18.38**
	XNLI	40.77	**43.00**
	XCOPA	55.55	**57.55**

Table 9: ***M2Lingual*** vs same size Aya-seeds (100K).

Benchmark	**No seeds** 157K	**25% seeds** 160K	**100% seeds** 175K
MT-EN	7.00	6.86	7.13
MT-FR	6.87 (6.80)	6.75 (6.79)	6.75 (6.81)
MT-IT	6.84	6.78	6.90
MT-JP	5.81	5.80	5.70
MT-ES	6.52	6.70	6.81
MT-DE	6.45	6.37	6.39
MT-NL	6.46	6.23	6.34
MT-Avg	6.57	6.55	6.54
XQUAD	71.49	72.96	74.53
TyDIQA	68.09	69.09	67.57
MLQA	58.57	61.62	62.40
XLSUM	9.38	9.45	10.42
MGSM	17.95	18.25	22.0

Table 10: Mistral-7B results with variable seeds on all benchmarks.

A.5 Importance of synthetic *Evols*

To assess whether the importance lies in the use of synthetic *Evol*s rather than simply increasing the amount of seed-like data, we sampled an additional 94.9K IR pairs from the Aya dataset and collection and replace it with synthetic generated *Evol*s. Results in 9 show that without synthetic *Evol*s the performance decreases, whereas having the same number of *Evol* IR pairs leads to higher performance especially in MT-Bench and MGSM by 1.1 and 3.38 points respectively.

Dataset	**25%**	**50%**	**100%**
MT-Avg	5.79	6.49	6.54
XQUAD	67.91	74.18	74.53
TyDIQA	63.67	67.93	67.57
MLQA	57.98	61.94	62.40
XLSUM	7.92	9.80	10.42
MGSM	16.05	17.55	22.00

Table 11: Mistral-7B performance results across benchmarks with different seed sizes used in figure 3.

A.6 *M2Lingual* performance without seed examples

Table 10 demonstrates the performance of our dataset (1) without seeds and (2) with 25% seed examples. Results show strong performance with multilingual MT bench without any seeds. It improves slightly compared to the last column that has all seed examples. The performance on other benchmark drops slightly but it still outperforms the evaluated baseline datasets in the paper.

A.7 Results with variable seed size

Finally, to show how performance changes as we scale synthetic data generation on more seed examples only, we ran 2 ablations where we (1) use only 25% of seed examples and use its synthesized data and (2) use 50% of the seeds. Figure 3 and Table 11 demonstrate that as we scale data synthesize by selecting more seed examples the performance increases across all benchmarks. Specifically, on an average we see 0.5 improvement in multilingual MT-bench, 5% in QA and MGSM and 2.50 in summarization.

A.8 Complete Results

Tables 12, 13, 14 and 15 shows the complete results comparing ***M2Lingual*** against all the baseline datasets, 4 base models across multilingual MT-Bench, question answering, summarization and classification tasks. Table 16 compares ***M2Lingual*** against top performing baseline on a smaller (Qwen1.8B) and a larger (LLama-2-13B) model.

QWEN-4B & Mistral-Instruct-7B results 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 14 and 15 respectively. Interestingly, the improvements from ***M2Lingual*** in Mistral-Instruct-7B over baseline datasets is consistently higher when compared to Mistral-7B-base (Table 13) 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.

A.9 Examples of Generated Evols and conversations

A.10 Prompt Taxonomy for Evol-instruct

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

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**
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**	**6.93**	**6.81**	6.74

Table 12: Multilingual MT-Bench results. Canadian French results are in MT-FR brackets. Best scores are in bold and dark green while 2nd best are in light green. *Seeds* are 15.1K seeds; *Seed + Evol* is additional *Evol* IR pairs. *Seed + Evol + MT* has additional multi-turn data.

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	13.65	42.28	57.64
	Seed + Evol	71.01	65.04	57.47	9.8	1.37	18.38	43.00	57.55
	Seed + Evol + MT (***M2Lingual***)	**74.53**	**67.57**	**62.40**	10.42	1.38	**22.00**	42.12	**59.55**
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

Table 13: Evaluations of LLaMA-3-8B-base & Mistral-7B-base in different tasks. Same notations as in Table 12

Model	Dataset	XQUAD F1	TyDiQA F1	MLQA F1	XLSUM ROUGE$_L$	 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
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**

Table 14: Evaluations of QWEN-4B & Mistral-Instruct-7B in different tasks and MT-Bench score averaged across languages. Please see table 15 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.

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
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

Table 15: MT-Bench evaluations in different languages for QWEN-4B and Mistral-Instruct-7B.

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**	**12.23**	**38.24**	**47.66**
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 16: Evaluations of QWEN-1.8B and LLaMa-2-13B for highlighting impact on different sized LLMs.

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 17: Conversation example from ***M2Lingual***

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 18: 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 19: 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 20: 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 21: 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 22: 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 Dont́ 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 23: 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 24: 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 25: 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 26: 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 27: 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 28: 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 29: 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 30: 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 31: 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 32: 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 33: 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 34: 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 35: 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 36: 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 37: 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 38: 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 39: 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 40: Openqa

A.11 Prompt Taxonomy for Multiturn Evol-instruct

B Licenses

We adhere to Apache 2.0 License from Aya Dataset and Aya Collection and Terms of Use for GPT-4 when constructing our ***M2Lingual*** dataset. We confirm that we bear the responsibility in the case of violation of rights and will take appropriate course of actions if needed. Our dataset is licensed through CC-by-NC-SA-4.0 license. The dataset will be hosted on HuggingFace datasets and maintained by the authors.

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 41: 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 5: Multiturn Prompt to GPT-4