LL-INSTRUCT: An Instruction-tuned model for English Language Proficiency Assessments

Anonymous ACL submission

Abstract

We present LL-INSTRUCT: an 8B instruction-tuned model designed to generate content for English Language Proficiency Assessments (ELPA). We leverage domain expertise to write seed instructions based on publicly available practice tests, which are then used by GPT-4 to generate 70K new instructions and explanations. GPT-4 is also used to validate its own generations, ensuring that the generated instructions are in-domain. Human evaluations show that a Llama-3 8B model fine-tuned on this dataset yields outputs comparable to GPT-3.5, with improved capability for generating explanations. A detailed error analysis highlights the strengths of our fine-tuned model, illustrating how it leads to improvements over standard out-of-the-box models on instructions related to English Language Assessments. To our knowledge, LL-INSTRUCT is the first instruction-tuned model designed specifically for ELPA generation.

1 Introduction

Instruction tuning—or multitask prompted finetuning—is an area in Natural Language Processing (NLP) that has led to state-of-the-art performance across a variety of tasks in recent years (Ouyang et al., 2022; Sanh et al., 2022; Wang et al., 2022; Peng et al., 2023). Broadly, this involves training a pre-trained Language Model (LM) using <INSTRUCTION, OUTPUT> pairs where the INSTRUCTION describes the task in natural language and OUTPUT is the desired outcome. As mentioned in Peng et al. (2023), current research focuses on two sub-areas: scaling instruction-tuning models using more training data and resources and/or utilizing large amounts of human-annotated data for training.

It is known that (a) gathering human-annotated data is expensive and time-consuming and (b) instruction-tuned models that are specialized for specific NLP tasks tend to struggle when applied to a diverse range of tasks (Zhang et al., 2023).

In response to these shortcomings, several recent models have been trained using a diverse set instructions generated via a semi-automated method. Wang et al. (2022) collected a small set of manually-written <INSTRUCTION, INPUT, OUTPUT> examples and then used the set of examples to prompt GPT-3 (Brown et al., 2020) to generate a larger set of more diverse instructions. The authors then fine-tuned GPT-3 using the generated tuples. This approach is named as SELF-INSTRUCT as the final model is trained on *self-generated* instructions.

Inspired by the aforementioned approach, we introduce **(L)anguage (L)earning INSTRUCT (henceforth, LL-INSTRUCT)**: an instruction-tuned model specifically designed for English Language Proficiency Assessments (ELPA) and other related applications in the language learning and test preparation domains. The language learning market is experiencing a significant growth with a projection of surpassing 100 billion USD in the coming years.[1] This coincides with the rise in the utilization of large pretrained LMs, with multiple organizations adopting automated content generation for ELPA such as Duolingo (Settles et al., 2020; Burstein et al., 2021) and Cambridge Assessment English (Galaczi, 2023).[2]

There are several aspects of LL-INSTRUCT that distinguish it from other instruction-tuned models.[3] Our **contributions** are as follows:

First, the authors manually write 130 seed instructions in the form of <INSTRUCTION, INPUT, OUTPUT> tuples based on publicly available test items from standard ELPA. These items are tailored

[1] https://www.meticulousresearch.com/pressrelease/792/language-learning-market-2030

[2] We use LLM and pretrained LMs interchangeably.

[3] Other instruction-tuned models include Alpaca (Taori et al., 2023), WizardLM (Xu et al., 2024), LLaMA-GPT4 (Peng et al., 2023), RoleLLM (Wang et al., 2024).

Instruction	Reply to the following email and express you can't attend the meeting due to a personal issue.
Input	We have scheduled a meeting at 4 PM tomorrow to discuss our project. Please confirm your availability.
Output	Thank you for the information. Unfortunately, I won't be able to attend the meeting tomorrow due to a personal issue. However, I'll make sure to catch up on the meeting notes.
Explanation	The response is polite and appropriate as it expresses the inability to attend the meeting due to a personal issue and includes an offer to catch up on what was discussed.
Instruction	Read the email. Create a question that asks about the scheduled date for the product launch.
Input	Email: Dear John, Just a quick reminder about the product launch happening next week. The team is keenly getting ready for it, and we appreciate all the hard work and dedication you've shown to make this successful. Best, Jane
Output	When is the product launch scheduled?
Explanation	The answer correctly specifies a question about the product launch date mentioned in the email.

Table 1: A few examples of instructions that were generated during the iterative prompting stage (Section 2.2). More examples can be found in Table 4 in Appendix A.2.

to assess skills such as reading, speaking, listening, and writing, which are crucial for language learning and testing (Section 2.1).

Second, an EXPLANATION for each output is generated. This feature can be highly valuable for test designers and test takers to understand the rationale behind the outputs. We use the seed tuples to iteratively prompt GPT-4 (OpenAI et al., 2024) to generate more data, i.e., 70K <INSTRUCTION, INPUT, OUTPUT, EXPLANATION> tuples (Section 2.2). Table 1 shows two instruction tuples generated during the iterative prompting stage.

Third, we use an LLM as a discriminator to reject any generated instructions that do not contribute to ELPA, such as, "What is the capital of Australia?" in a separate evaluation stage, similar to Chiang and Lee (2023) and Gao et al. (2023). The data quality is further verified through human annotation (Section 3).

Fourth, we fine-tune three Llama-3 8B (Meta, 2024) models with different dataset sizes: 17K, 50K, and 70K (Section 4). Contrary to previous work suggesting that the effect of data size on model performance plateaus after 16K instructions (Wang et al., 2022), we find marked improvement from 50K to 70K instructions.

Finally, we conduct a comprehensive human evaluation for 200 unseen instructions. Several pre-trained models are evaluated alongside the fine-tuned models: Dolly-2 8B (Conover et al., 2023), Mistral 7B (Jiang et al., 2023), Llama-3 8B (Meta, 2024), and GPT-3.5.

The quantitative results are discussed in Section 5.1. We find that models tend not to work out-of-the-box for this task and require domain adaptation. The models we train through **S**upervised **F**ine-**T**uning (SFT) are comparable, but SFT-70K produced most outputs that are valid and ready for use in ELPA. Specifically, both SFT-70K and GPT-3.5 produce over 60% valid and ready outputs exceeding Dolly-2, Mistral, and Llama-3 base version. SFT-70K also generated the most usable explanations while rates were much lower for non-SFT models (SFT-70K: 80.5%, GPT-3.5: 42%). In Section 5.2 we describe a detailed qualitative error analysis that illustrates how the SFT process leads to improvements over GPT-3.5 and other open source models. All datasets and models will be released upon acceptance.

2 Dataset Creation

In this section we describe in detail our approach of collecting training data to build the LL-INSTRUCT model. This includes: (1) writing seed instructions, (2) generating diverse instructions with an LLM, and (3) filtering the generated instructions. The full flow is shown in Figure 1.

2.1 Writing Seed Instructions

The authors leverage domain expertise from years of working with standard ELPA to convert publicly available ELPA items into seed instructions, distinguishing our approach from generalized datasets that aim to cover a broad range of knowledge such as those used to train Alpaca (Taori et al., 2023) and Dolly (Conover et al., 2023).[4]

An ELPA item, often comprising a stimulus (e.g., source text) and multiple-choice questions, is typically divided into multiple instructions, as many items include sub-items. Each instruction may fo-

[4]In the assessment domain, an "item" is a term commonly used to denote a stimulus accompanied by a question/answer set. Many companies conducting ELPAs globally provide sample tests online.

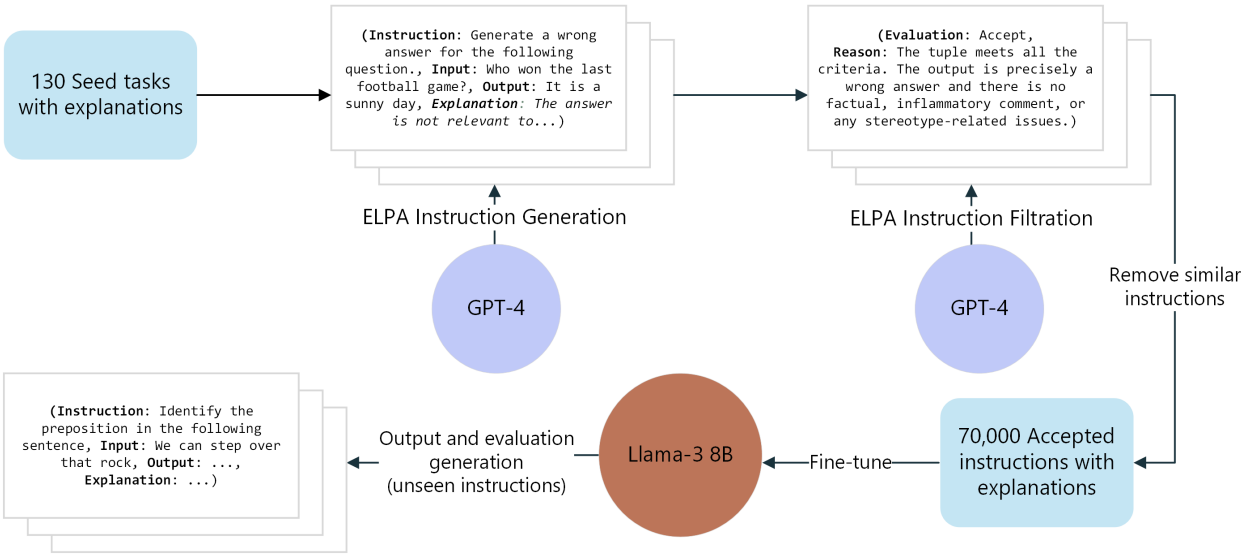

Figure 1: Overview of our proposed instruction data generation and filtration strategies. We begin by writing seed instructions (Section 2.1), followed by ELPA instruction generation (Section 2.2) and then ELPA instruction filtration (Section 2.3). Accepted instructions are used for fine-tuning the the Llama-3 8B model.

cus on a specific part, such as the question, correct answer, or incorrect options, or cover multiple parts. This approach caters to ELPA test designers, who can use several <INSTRUCTION, INPUT, OUTPUT> tuples to construct a complete ELPA item.

Table 2 illustrates this process. The first column, "Original Item", shows a sample ELPA item (available online) where the test taker responds to "Who is your favorite tennis player?" In the second column, "Seed Instruction", we created two <INSTRUCTION, INPUT, OUTPUT> tuples from it.

Given the best response "I don't like any sports" is an *indirect* answer to the question, we create a seed instruction (e.g., "write an *indirect* response ...") to reflect that. Likewise, we create the second instruction that focuses on generating *three wrong answer*s, commonly known as the distractors that are designed to divert test takers from selecting the correct answer.

Each tuple also includes an EXPLANATION, which proves helpful in understanding the reasoning behind each output. Seed explanations (Table 2) are written by the authors of the paper. It is not always necessary to generate more than one seed instruction from a test item; the goal is to create a variety of seed instructions that capture all the different components we wish to generate using the LL-INSTRUCT model in the end.

In total 130 seed instructions were created from standard ELPA. To increase variation, input text is sometimes included in the INSTRUCTION and other times in the INPUT section. We also adopt linguistic variations in the instructions by using different phrases for the same type of tasks, like "generate the answers," "create a sentence," "write the answer," and so forth to mitigate the decline of linguistic diversity found in LMs trained on synthetic text (Guo et al., 2024).

2.2 ELPA Instruction Generation

New instructions are generated via a bootstrapping method that passes few-shot examples to GPT-4, closely following Wang et al. (2022) and other synthetic data generation approaches described in Long et al. (2024). In each step, four <INSTRUCTION, INPUT, OUTPUT, EXPLANATION> instruction tuples are included in the prompt: three seed tuples and one model-generated tuple. To promote diversity, we split the 130 seed tuples into two categories: **(a) short tasks** (e.g., grammar correction, convert a passive sentence into active, etc.) and **(b) long tasks** (e.g., write an email, write a short conversation, etc.) and then randomly choose either two short tasks and one long task or vice-versa in the prompt. Then, GPT-4 is prompted to generate ten new tuples, corresponding to ten new instructions.

Refer to Section A.1.1 in the Appendix section for the prompt template. We have included a few requirements in the prompts that are intended to guide the model. For example, we ask that the new instructions be relevant to ELPA and not involve

Original Item	Seed Instruction
Choose the best response. Statement: who is your favorite tennis player? Answers: [‘play the song again’, ‘my favorite color is green’, ‘that sounds like fun’, **‘I don’t like any sports’**]	INSTRUCTION: Write an indirect response to the following question. INPUT: who is your favorite tennis player? OUTPUT: I don’t like any sports. EXPLANATION: Reply indirectly addresses the question by stating a general disinterest in sports rather than specifically identifying a favorite tennis player. INSTRUCTION: Write three wrong answers to the question. INPUT: who is your favorite tennis player? OUTPUT: [‘play the song again’, ‘my favorite color is green’, ‘that sounds like fun’] EXPLANATION: The given answers are either unrelated (e.g. “play the song again”), changing the subject (e.g. talking about a favorite color), or expressing a general sentiment (e.g. commenting on an activity).

Table 2: Original ELPA item (first column) and extracted seed instructions (second column) from the items. The **bold** selection presents the correct choice. Original items are lightly edited for anonymity.

generating code or solving arithmetic problems.

2.3 ELPA Instruction Filtration

Evaluation of LLM outputs guided by another LLM has been shown to be effective (Chiang and Lee, 2023). Thus, to remove factual data from the **ELPA Instruction Generation** round, we again prompt GPT-4 using few-shot examples. Although in the generation round we ask GPT-4 to exclude factual data such as “What is the capital of Australia?”, the model does not always adhere.

We create a new filtering prompt (refer to Section A.1.2 in the Appendix for the complete template) that includes examples of both factual and non-factual tasks, the latter being what we want to keep in the dataset. In a pilot test we found that approximately 7% of tasks from the bootstrapping step (Section 2.2) are flagged as factual and removed.

The filtering prompt is used along with the ROUGE-L similarity metric to remove unwanted instructions during the generation process, resulting in a set of diverse, non-factual instructions. A newly generated instruction is discarded if it is labeled as factual or if it has a similarity score ≥ 0.75 when compared to an instruction already in the dataset (this value was set empirically after tuning). After each batch of ten instructions is generated, we apply filtering to that batch. This continues until the desired number of 70K instructions is reached.

3 Evaluation of Data Quality

This section focuses on the evaluating the quality of the generated LL-INSTRUCT data. Our work differs from Wang et al. (2022) and related studies in that we conduct a large-scale evaluation to ensure the instructions’ relevance and suitability for our domain. For more examples of generated instructions and inputs, see Section A.5.2 in the Appendix.

We randomly selected 250 generated instruction tuples and carried out the evaluation in two stages. First, we classify the instruction tuples by language **category** (e.g., grammar, semantic, etc.) and language **skills** (e.g., speaking, writing) to illustrate the types of instructions included in the dataset (Section 3.1). Importantly, the language category or skill could be identified in all 250 instructions, confirming that the entire sample was in-domain.

Next, we further focus into specific aspects of the instruction tuples, such as, output correctness, quality of explanation, etc. (Section 3.2).

3.1 First Evaluation Task: Language Category and Skills

The language **category** of an instruction defines the type of linguistic knowledge the <INSTRUCTION, INPUT, OUTPUT> tuple and the resulting ELPA item assesses, such as grammar, vocabulary, semantics, pragmatics, and prose (e.g., prose writing).[5] Pragmatic and figurative tasks also appear frequently (e.g., “identify informal words,” “write a simile about an everyday item”). Besides these, interestingly, tasks specific to language learning assessments, like Build a Sentence (abbreviated as BaS in Figure 2) also appear. Likewise, instructions

[5] Items in standardized ELPAs include specific constructs related to grammar, semantics, and pragmatics, typically authored by assessment developers.

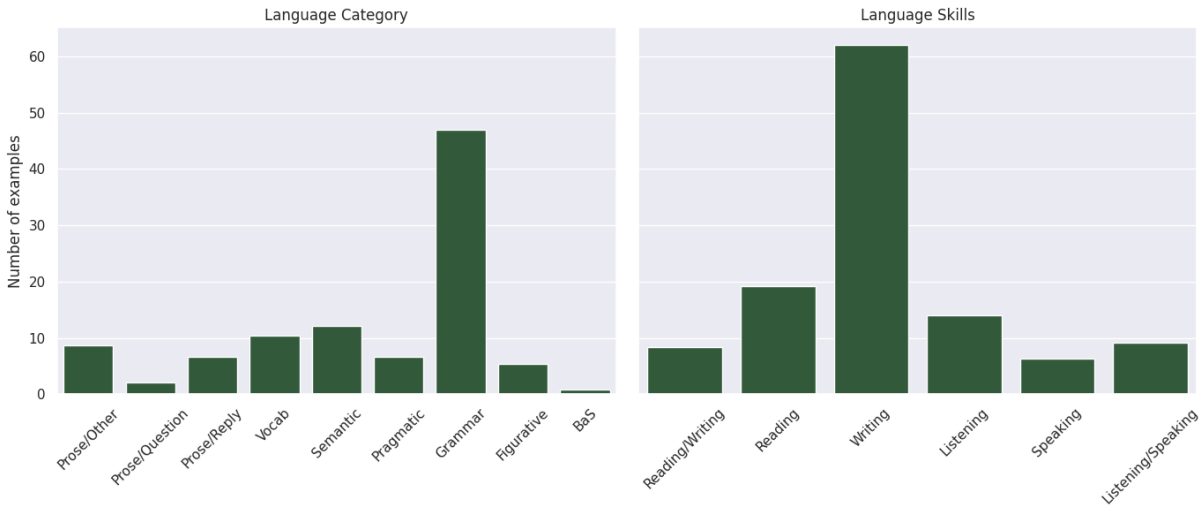

Figure 2: Percentage of LL-INSTRUCT data by language categories (a) and skills (b).

are grouped by **skill**: reading, writing, speaking, or listening.[6] A pilot annotation task was conducted to establish the main categories, followed by annotation of the full dataset by the two authors. The authors' domain expertise comes from working on automated ELPA item generation for several years. Figure 2 shows the identified categories and skills.

Language Category In terms of the observed categories, grammar-related instructions are common (e.g., "rewrite the question in reported speech"), alongside a variety of prose-writing tasks categorized as Prose/Question ("write questions based on the passage"), Prose/Reply ("write a reply to the dialog"), and Prose/Other ("write an opinionated argument on topic x"). Krippendorff's α (Krippendorff, 2011) among the two authors was 0.79, indicating substantial agreement.

Language Skills Regarding the types of language skills, we observe that the majority can be classified as writing tasks. This includes most grammar and semantic instruction categories, in addition to the clearly defined prose category. We also identified some instructions as both reading/writing tasks (e.g. "read the following sentence and paraphrase it in the past simple tense", "read the statement and suggest an alternative word ..."), which can be seen as two combined tasks. Finally, we also notice a few instructions that can be categorized as listening, speaking, or speaking/listening. Krippendorff's α (Krippendorff, 2011) among the two authors was 0.83, indicating substantial agreement.

[6]This four skills approach is widely adopted by ELPA and the language teaching community. While Powers (2010) advocates assessing these skills individually, Hinkel (2010) notes it is possible to integrate them in pedagogy.

3.2 Second Evaluation Task: Instruction Quality

Here, we focus onto the following aspects of the instructions:

- **Validity**: whether the example is valid and ready to appear in an English language assessment. We provide three options: *valid and ready* for assessment, only *valid* (i.e., needs some editing), and *invalid*.
- **Instruction type**: whether the instruction is *factual* or *not factual*.
- **Input faithfulness**: whether the input *matches* or *does not match* the instruction.
- **Output correctness**: whether the output is *correct* (based on the instruction) or not.
- **Explanation quality**: does the explanation justify the output? We provide four options: *yes*, *weak yes*, *weak no*, and *no*.

We recruited ten expert annotators with backgrounds in linguistics and computer science, and work experience in educational technology. Each pair evaluated 50 instruction tuples (see Section A.3 for guidelines). Krippendorff's α (Krippendorff, 2011) was calculated for each aspect across pairs, with average scores as follows: **Validity** (0.49, moderate agreement), **Instruction type** (0.93, almost perfect agreement), **Input faithfulness** (0.67, substantial agreement), **Output correctness** (0.78, substantial agreement), and **Explanation quality** (0.52, moderate agreement).

We further focus into two aspects where agreement was lower. For **Validity**, most disagreements

concern labeling instructions as *valid and ready* versus *valid*. For example, some annotators critiqued instructions like "identify the tense/voice/verb type.. . ." saying that an assessment would provide a list of possible options. Open-ended tasks like "provide a synonym/alternate ending. . . " were also flagged as overly ambiguous. For **Explanation quality**, disagreements were often around whether the explanation was sufficient (*yes*) or required human editing (*weak yes*).

4 Experimental Details

The experiment involves Supervised Fine-Tuning (SFT) of a Llama-3 8B model. The design choice to use a small 8B model is driven by two primary motivations: (a) to evaluate how effective a small SFT model can be for language learning applications, and (b) to ensure fast inference and moderate GPU requirements, thereby lowering the barrier to trying these model(s).

Fine-tuning A Llama-3 8B model was fine-tuned on subsets of the LL-INSTRUCT data of size 17K, 50K, and 70K.[7] Each <INSTRUCTION, INPUT, OUTPUT, EXPLANATION> tuple was joined into one example using the following template:

> Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: INSTRUCTION ### Input: INPUT ### Output: OUTPUT### Explanation: EXPLANATION

HuggingFace (Wolf, 2019) is used to perform the SFT.[8] We use one Amazon AWS g5.12xlarge instance (NVIDIA 4 GPUs) for training.

Inference on Test Dataset To evaluate the performance of the SFT-17K, SFT-50K, and SFT-70K models, we compare them to the following: Llama-3 8B base (Meta, 2024), GPT-3.5 (Brown et al., 2020), Mistral 7B (Jiang et al., 2023), and Dolly-2 8B (Conover et al., 2023). The Mistral and Dolly models were chosen due to their similar size and their state-of-the-art performance on a variety of NLP tasks.[9] In addition, we selected GPT-3.5 as we wanted to compare to a larger model. We are also curious as to how a model in the same family as the model we are distilling knowledge from (GPT-4) will compare. Note that we compare against zero-shot, non-finetuned models. The prompt for inference is similar to the one used for fine tuning, except we start generating at the output:

> Write the output by following the instruction and the input, and then include an explanation for why the output is appropriate given instruction and input. Include a separator token '###' before the explanation.
> ### Instruction: INSTRUCTION ### Input: INPUT ### Output:

We selected an unseen batch of 200 instructions for the comparison, where the instructions are ranged over diverse tasks such as grammar, figurative language and prose.

5 Human Evaluation of Model Performance

For each test instruction, the authors (who have worked with ELPA for several years) jointly evaluated the output from each of seven models, i.e., a total of 200x7, 1400 outputs. The model was not hidden during evaluation. We re-use the rubric that was used to evaluate the quality of LL-INSTRUCT in Section 3.2. We assessed dimensions of **Validity**, **Output correctness**, and **Quality of explanation**, but omitted **Instruction type** and **Input faithfulness** dimensions due to the non-factual nature of almost all instructions and the lack of dedicated input entries in the test set.

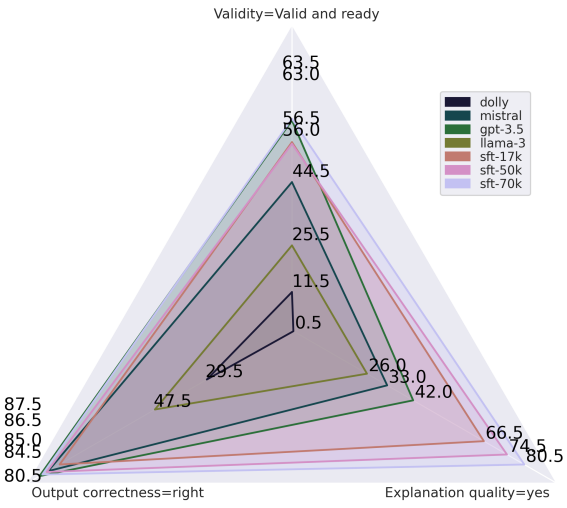

Figure 3: Comparison of human evaluation results across all seven models on three dimensions (Validity, Output correctness, and Explanation quality).

[7] We chose the 17K and 50K partitions randomly from the total of 70K instruction tuples that are generated.

[8] Parameter specifications: PEFT with LoRA setting (r=16; α=32), Adam optimizer (default learning rate) for 5 epochs with a cosine learning rate scheduler (Loshchilov and Hutter, 2017).

[9] Vicuna (Chiang et al., 2023) was also evaluated and found to perform similar to Mistral.

	Dolly-2	Mistral	GPT-3.5	Llama-3	SFT-17K	SFT-50K	SFT-70K
Validity							
valid and ready	11.50	44.50	**63.00**	25.50	56.50	56.00	**63.50**
valid	16.50	41.50	23.50	22.50	24.00	29.00	22.50
invalid	72.00	14.00	13.50	52.00	19.50	15.00	14.00
Output correctness							
right	29.50	84.50	**87.50**	47.50	80.50	85.00	**86.50**
wrong	70.50	15.50	12.50	52.50	19.50	15.00	13.50
Explanation quality							
yes	0.50	33.00	42.00	26.00	66.50	74.50	**80.50**
weak yes	1.00	32.50	9.00	17.00	9.50	7.00	5.50
weak no	3.00	5.00	1.00	3.50	1.00	0.50	0.00
no	95.50	29.50	48.00	53.50	23.00	18.00	14.00

Table 3: Results from human evaluation of model performance. Percentage of a models' output along dimensions of: Validity, Output correctness, and Explanation quality. Best-performing model(s) for each dimension are **bolded**.

5.1 Model Comparison

An overview of how the models fared on each dimension can be seen in Figure 3. Table 3 summarizes the results. Additional tables that show the win-rate and tie-rate between each model can be found in Section A.4 in the Appendix.

Validity SFT-70K and GPT-3.5 have the highest number of *valid and ready* generations where SFT-70K is marginally better (63.5% vs. 63%, see Table 3) followed by SFT-50K and SFT-17K almost equally. In contrast, Mistral, Llama-3, and Dolly-2 have fewer than 50% valid and ready generations, with Mistral leading at 45%, followed by Llama-3 at 26%, and Dolly-2 at 12%.

Output Correctness GPT-3.5 has the highest number of *right* generations (87.5%), followed closely by SFT-70K (86.5%), whereas the two remaining SFT models, SFT-50K and SFT-17K perform similar to Mistral (all between 80-85%). The remaining two models, Llama-3 (base model) and Dolly-2 have fewer than 50% right generations.

Explanation Quality The explanations are rated on a four-category scale. SFT-70K has the most explanations in the *yes* category (80.5%), followed by SFT-50K (74.5%), then SFT-17K (66.5%). Perhaps the SFT models showcase their usefulness the most for this aspect, given, the closest best explanations are from GPT-3.5 (42%), which is 38.5% lower than SFT-70K model's performance.

5.2 Qualitative Error Analysis

We conducted a thorough error analysis of all the model outputs and highlighted specific characteristics here. Refer to Section A.5.2 in the Appendix for all outputs.

Verbose outputs and explanations Most often for GPT-3.5, Mistral, and Llama-3, while the output and explanation match the specification, an excessive amount of words and description is used.

For the same instruction GPT-3.5 produced a 212-word email, while SFT models created emails around 100 words. Without specified word limits, GPT-3.5 often wrote very long responses, sometimes reaching 250-300 words. Llama-3 (base) and Mistral also frequently created longer responses.

Our evaluation overlooks verbosity unless it makes the output ineligible, since it does not impact Validity or Output Correctness.

Explanations are often missing We find that often the explanations are missing. For instance, we notice that explanations are missing from 95% of Dolly-2 generations, 55% of Mistral generations, 51% of GPT-3.5 generations, and 10% of Llama-3 generations. On the contrary, only less than 1% of any fine-tuned model (SFT-17K, SFT-50K, and SFT-70K) is missing the explanation.

Formatting errors of outputs Formatting errors in the generation can hinder the full automation of ELPA item generation. Common formatting issues include a numbered list being returned when only one item is requested, typically by the base Llama-3 model. Other problems involve the separator token ### (between the OUTPUT and EXPLANATION) being misplaced, repetition of the instruction in the output, and so on. However, SFT models exhibit less frequently such formatting errors.

Outputs are often in the proximity but do not follow the instruction exactly We often notice errors where the models' interpretation of the instruction is close, yet it does not adhere to the request by missing some part of the instruction or simply adding extra information (hallucination).

Consider the following instruction:

> INSTRUCTION: Paraphrase the sentence.
> INPUT: Nobody knew how much time she spent training for the Olympic Games.

The Dolly-2 generation contains a close paraphrase but adds extra hallucinated information imagining the content is regarding a freestyle skier. Likewise, Llama-3b base produces: *"She trained a lot for the Olympic Games"* which fails to fully convey the input's meaning. Interestingly, the SFT models excel at accurately capturing this precise instruction.

Tasks involving figurative language are difficult Grammar, vocabulary, and simpler prose tasks tend to be cases where most models produce valid output. On the other hand, figurative language tends to be difficult for most models. Even the SFT-50K model sometime misses generating such content that includes a figurative type, e.g., an idiom (See Example A.5.4 in the Appendix).

6 Related Work

Recent studies have demonstrated that LMs can effectively follow language instructions when fine-tuned using human-annotated datasets that pair instructions with outputs (Weller et al., 2020; Sanh et al., 2022; Peng et al., 2023). To address the reliance (bottleneck) on large-scale human annotations, researchers like Ouyang et al. (2022) have developed general-purpose LMs for diverse instructions. Our research closely aligns with Wang et al. (2022)'s self-instruct method (i.e., an iterative method for creating new instructions and outputs to enhance fine-tuning); this method has since been adopted by many models.[10] Our work fills a research gap on instruction-tuned models specialized in the language learning and assessment domain.

Similar to our approach, *Humpback* (Li et al., 2023) curated seed instructions to generate new ones.[11] A key distinction is that *Humpback* instructions are drawn from existing web corpora, whereas we generate all components of the <INSTRUCTION, INPUT, OUTPUT> tuple. Although our work also relates to the self-training literature, our approach is different. Our instructions exhibit wide diversity across various instruction types while keeping within the ELPA domain. Lastly, our research aligns with the concept of distillation (Hinton et al., 2015), as we extract new instructions from a teacher model (in this case, GPT-4). We also employ a language model as a discriminator to eliminate factual inaccuracies and non-ELPA instructions (Chiang and Lee, 2023; Gao et al., 2023).

7 Conclusion and Future Work

We compiled instruction seed data consisting of <INSTRUCTION, INPUT, OUTPUT, EXPLANATION> tuples designed for item generation in ELPA. Using these, we prompted GPT-4 to generate a much larger dataset of instruction tuples. Subsequently, we fine-tuned Llama-3 model using with different partitions (17K, 50K, and 70K) of the data.

We compare the performance of the fine-tuned models against various LM baselines including Dolly-2, Mistral, Llama-3 base model, and GPT-3.5. The fine-tuned versions consistently demonstrate superior performance in terms of output validity, correctness, and explanation quality (Section 2.3). In contrast to prior work suggesting performance gains from dataset size plateaus at around 16K instructions (Wang et al., 2022), we find that our model fine-tuned on 70K compared to 50K instructions shows a marked improvement.

Our detailed error analysis identified common issues across the models. While the fine-tuned models produced approximately 60% of test-ready outputs, about 20-30% require manual adjustments by subject matter experts. This suggests that a combined human-AI approach would be most effective for advancing ELPA task designs.

For future work, we plan to improve our SFT model by aligning with human preference, e.g., DPO (Rafailov et al., 2023). We also plan to align the trained models to specific attributes by post-hoc merging of parameters (similar to Jang et al. (2023)).

8 Ethics

The risks and harms of language models are well-documented. Bender et al. (2021) provides an overview, including: environmental and financial cost; unfathomable training data leading to encoded

[10] Such as Alpaca (Taori et al., 2023), WizardLM (Xu et al., 2024), LLaMA-GPT4 (Peng et al., 2023), RoleLLM (Wang et al., 2024), to name a few.

[11] Other synthetic data generation approaches are surveyed in Long et al. (2024).

biases that reflect the dominant/hegemonic view; coherent output being mistaken as true knowledge.

This work uses GPT-4 (1.76 trillion parameters) for dataset generation. Apart from increased water consumption and carbon emissions (Strubell et al., 2020; George et al., 2023) when using a larger model, there is the risk of including harmful biases and misinformation in both the training data and in the fine-tuned models. The data was spot-checked and filtered using another LLM to remove factual data. To mitigate bias and fairness issues, we recommend adding additional checks, such as those described in Stowe et al. (2024), and involving human reviewers before rolling out machine-generated content to learners or test takers.

We hope to show that a smaller model (i.e., a model that consumes less resources) can achieve the same performance as that of a larger model when training data is available. While smaller models are more accessible, they remain difficult to access in resource-limited environments where GPU compute is rare or expensive.

9 Limitation

The experiments were conducted for Llama-3 8B, and it is uncertain whether the findings will generalize to other models. The human evaluation of model performance was completed by the authors, who were also designed and conducted the experiment. As there may be unconscious biases on part of the authors, the dataset and annotations will be released upon acceptance.

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jill Burstein, Geoffrey T LaFlair, Antony John Kunnan, and Alina A von Davier. 2021. A theoretical assessment ecosystem for a digital-first assessment—the duolingo english test. *DRR-21-04*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. 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)*, 2(3):6.

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.

Evelina Galaczi. 2023. English language education in the era of generative ai: Our perspective.

Jiahui Gao, Renjie Pi, LIN Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023. Self-guided noise-free data generation for efficient zero-shot learning. In *International Conference on Learning Representations*.

A Shaji George, AS Hovan George, and AS Gabrio Martin. 2023. The environmental impact of ai: a case study of water consumption by chat gpt. *Partners Universal International Innovation Journal*, 1(2):97–104.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

Eli Hinkel. 2010. Integrating the four skills: Current and historical perspectives.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

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*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. *University of Pennsylvania ScholarlyCommons*.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

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, Mohammad 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, Ben 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, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha 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, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz 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 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 Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde 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 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 Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll 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. 2024. Gpt-4 technical report.

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

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Donald E Powers. 2010. The case for a comprehensive, four-skills assessment of english-language proficiency. *R & D Connections*, 14:1–12.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Kevin Stowe, Benny Longwill, Alyssa Francis, Tatsuya Aoyama, Debanjan Ghosh, and Swapna Somasundaran. 2024. Identifying fairness issues in automatically generated testing content. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 232–250, Mexico City, Mexico. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696.

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`.

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

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models.

Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E. Peters. 2020. Learning from task descriptions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1361–1375, Online. Association for Computational Linguistics.

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

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. 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*.

A Appendix

A.1 Prompt templates

A.1.1 Prompt template to generate ELPA instructions

Generation prompt template

You are asked to come up with a set of 15 task instructions in English. These instructions should be useful for language learners of English. These task instructions will be given to a GPT model and we will evaluate the GPT model for completing the instructions. Separate each instruction using "###".
Here are the requirements:
1. The type of instructions should be similar and related to the instructions in the prompt.
2. These instructions should be related English language learning, such as grammars, semantics, pragmatics, etc.
3. Please don't write instructions to write a code or program or answer a mathematical question.
4. Please avoid generating factual instructions that ask specific questions on history, geography, politics, or science.
5. The instructions should not contain racist, sexist, toxic, or otherwise potentially offensive language.
6. Not all instructions require input. For example, when an instruction asks "did you have lunch yet", it is not necessary to provide a specific context. In this case, we simply put "<noinput>" in the input field.
7. The output should be an appropriate response to the instruction and the input.
List of 15 tasks:

// Here we insert a 3 seed instructions and 1 model-generated example.
1. Instruction: ...
1. Input: ...
1. Output: ...
1. Explanation: ...
###
...
###
5. Instruction: ...
5. Input: ...
5. Output: ...
5. Explanation: ...
###
6. Instruction:

A.1.2 Prompt template to filter ELPA instructions

Filtration prompt template

Given the following tuples of <instruction,input,output> your task is to evaluate the quality of the tuple(s) and accept or reject them based on the following requirements. Separate each example using "###".
Here are the requirements:
1. The <instruction,input,output> tuples are used for language learning in English.
2. The output should not contain any verifiable factual information related to science, geography, history, business etc.
3. The output should not contain any code, program, or mathematical formula.
4. The output should be free of racist, sexist, toxic or otherwise potentially offensive language and imagery.
5. The instruction should not contain inflammatory, highly controversial or upsetting topics.
6. The output should not contain language or symbols that reinforce stereotypes.
7. Return the full tuple <instruction,input,output> with Evaluation and Reason.

See the following examples.
1. Instruction: Reply the statement with a proper answer.
Input: we are all going to the movie at 7pm.
Output: Great, my work finishes at 6pm so perhaps I can join too.
Evaluation: Accept.
Reason: The instruction, input, and the output does not contain any factual information, code, or inflammatory comment.
###
2. Instruction: Write an indirect answer to the question.
Input: who is your favorite soccer player?
Output: I loathe soccer because it is not a manly sport.
Evaluation: Reject.
Reason: The output seems toxic ("not a manly sport") and gender biased.
###
3. Instruction: This is an email written by a customer to a customer support team. Please give me a question that asks about the main idea.
Input: Email: Hello, Thanks for sending my order #3397—it arrived this morning. Unfortunately, the paint was not the one I had asked for. I had selected color SP 944 but received SP 945 (Ocean Waves). They appear right next to each other on your Web site, so the two may have been confused at your end. Could you send me the correct paint, along with additional samples that are close in color to SP 722? Thank you, Arun Phan
Output: What problem does Mr. Phan mention in his e-mail?
Evaluation: Accept.
Reason: The instruction, input, and the output does not contain any factual information, code, or inflammatory comment.
###
4. Instruction: Write a 6-turn exchange between 3 people (Person-1, Person-2, and Person-3). They all work at the same company, and discuss thoughts on which division will end up occupying the space.
Input: no-input.
Output:
"Person-1: Have you two taken a look at the progress they've made upstairs on the office expansion? It looks great!
Person-2: I know! I can't believe it! And the offices up there have amazing views of the city.
Person-3: I wonder which division will move up there when it's finished.
Person-2: I heard it's the research department.
Person-1: Ah, because the CEO is biased towards the department. In fact the CEO hired her husband to lead a new project inside research.
Person-1: I think you're right, there!"
Evaluation: Reject.
Reason: The output seems toxic. The conversation is not suitable for a workplace environment.
###
5. Instruction: What is the capital of India?
Input: no-input.
Output: New Delhi.
Evaluation: Reject.
Reason: The output is factual.

A.2 More examples of generated data

Instruction	Convert the following informal sentence to a formal one.
Input	Can you give me a hand with this report?
Output	Could you kindly assist me with the completion of this report?
Explanation	The output changes the informal request "Can you give me a hand" to a more formal and polite request "Could you kindly assist me".
Instruction	Write a sentence using the given idiom.
Input	"at the drop of a hat"
Output	John was known to change plans at the drop of a hat, often leaving his friends confused.
Explanation	The phrase "at the drop of a hat" is skillfully incorporated into a sentence, illustrating the idiom's meaning by describing a scenario where John readily changes his plans.
Instruction	Convert this assertive sentence into an interrogative one without changing its meaning.
Input	"You know the deadline for the project is next week."
Output	"Do you know the deadline for the project is next week?"
Explanation	The output is a question that keeps the meaning of the original assertive sentence by asking about the knowledge of the project deadline being next week.
Instruction	Make a formal apology to a customer.
Input	I am upset that the product I was sent doesn't match the product I was shown.
Output	Dear Customer, We greatly apologize for the guidance and inconvenience that you received in relation to the product you ordered. This is certainly not the level of service we strive to provide, and we want to make it right. We are happy to issue a replacement with the correct product at no extra cost. Again, our sincerest apologies for the blunder. We promise to make this right.
Explanation	The reply is a formal, gracious apology that admits fault on the company's behalf, provides a solution to the issue, and reiterates the commitment to good service.
Instruction	Formulate a question that could be answered with the statement provided.
Input	I haven't completed it yet, but I'm about halfway done.
Output	How far along are you with the project?
Explanation	The question asks about the progress of an unspecified task, which is then addressed in the given statement, indicating a progress of being approximately halfway done.
Instruction	Build a 6-turn conversation around a topic. In the conversation, use the quote, "Actions speak louder than words.
Input	noinput
Output	Person 1: Do you think apologizing is enough when someone hurts you? Person 2: Well, apologies matter but only if they are sincere. You know what they say, "Actions speak louder than words." Person 1: True, we often forget the impact of our actions while relying too much on our words. Person 2: Exactly, it's easy to say sorry, but you can tell it's sincere when the actions change. Person 1: I totally agree with you. It's important to see that change. Person 2: Absolutely, it's all about the actions in the end
Explanation	The conversation discusses the importance of actions along with words, effectively integrates the quote and revolves around the topic of apology and change in behavior which validates the saying "Actions speak louder than words

Table 4: The additional examples include instructions from these categories (see Section 3.1 for more information): pragmatic, figurative, grammar, prose, prose/question, and prose/dialog.

A.3 Annotation guideline for evaluating the instruction quality

Instruction-tuning models have become one of the cornerstones of recent NLP research. We have automatically generated instructions using GPT-4 following some specific input prompts. Our goal is to generate enough valid instructions that can be used to finetune a new instruction model.

The annotation task here is to evaluate the quality of the instructions. With each instruction, we also have an input, an output, and in some cases, an explanation. See the following example,

Instruction: Reply the statement with a proper answer.
Input: we are all going to the movie at 7pm.
Output: Great, my work finishes at 6pm so perhaps I can join too.
Explanation: The reply acknowledges the information given in the context and suggests a potential plan for joining the group at the movie at 7pm.

We will be evaluating each component (instruction, input, output, explanation) based on five different aspects. Your task will be simply selecting from the dropdown menus (items will be provided on Excel sheets). In most cases, the options are self-explanatory.

Aspect	Instructions
Validity	1. Is the instruction valid as well as can be asked in a language learning assessment? E.g. the above example ("reply the statement . . . ") falls into this category. Select "valid and ready". 2. Is the instruction seeming valid but not the type of instruction we use in a language learning assessment? Such as an instruction like – "who is the captain of the USA soccer team" or "what is the capital of Canada". For such instructions, select "valid". 3. Invalid instruction, that is illogical, does not follow common-sense, or just a garbage statement. Select "invalid".
Instruction type	1. Although we prompted the LLM not to generate instructions that are factual or based on real world events (such as a question asking the capital of a country) sometime the LLM still does generate such instructions. We will categorize these as "factual". Note, instructions like "write a program/code" or "draw an image" also belong to this category. Select "factual". 2. Non-factual instructions. These are ones that relate to language learning, typical of the instructions that we want the model to train with (e.g., reply a statement/grammar check/write an idiom using x/) etc. Select "non factual".
Input faithfulness	1. If the input follows the instruction perfectly and matches what has been instructed. For instance, for the above example, the input is a proper statement that could be replied. Select "matches". 2. If the input does not follow the instruction we call it not-matched. Select "does not match".
Output correctness	1. Is the output correct for the instruction and input pair? Select "right". 2. Otherwise, select "wrong".
Quality of explanation	1. Yes, the explanation correctly explains the output (select "yes"). 2. The explanation is right but could be better (select "weak yes"). 3. The explanation is not totally correct (select "weak no"). 4. The explanation is wrong (select "no").

A.4 Model performance results

	Dolly	Mistral	GPT-3.5	Llama	SFT-17K	SFT-50K	SFT-70K
Dolly		8.50 4.50 1.50	4.00 2.00 2.00	15.50 14.50 3.50	7.50 4.00 1.00	6.00 3.00 0.00	4.50 2.00 0.00
Mistral	71.00 59.50 69.50		20.00 9.00 35.00	54.00 42.00 44.50	25.50 14.50 16.50	23.00 11.00 12.00	24.00 10.50 13.00
GPT-3.5	72.50 60.00 51.00	34.50 12.00 29.50		56.50 43.00 37.00	25.50 13.50 14.50	27.50 11.50 11.50	22.50 11.00 12.00
Llama	38.00 32.50 46.50	15.50 5.00 21.00	8.00 3.00 23.00		12.00 6.50 8.50	10.50 4.50 5.00	7.00 4.50 5.00
SFT-17K	68.50 55.00 76.50	33.00 10.50 44.50	18.50 6.50 39.50	55.00 39.50 54.50		24.00 10.50 12.50	18.00 8.50 10.50
SFT-50K	72.00 58.50 81.50	32.50 11.50 51.50	22.50 9.00 46.50	54.50 42.00 58.50	26.00 15.00 21.00		15.50 8.00 11.00
SFT-70K	70.50 59.00 85.50	39.00 12.50 56.00	23.50 10.00 50.00	56.50 43.50 60.50	26.00 14.50 25.00	22.50 9.50 16.50	

Table 5: The table represents the **win-rate** of each model on the y-axis when compared to the model on the x-axis. Validity in dark gray, Output correctness in olive, and Quality of explanation in magenta.

	Dolly	Mistral	GPT-3.5	Llama	SFT-17K	SFT-50K	SFT-70K
Dolly		20.50 36.00 29.00	23.50 38.00 47.00	46.50 53.00 50.00	24.00 41.00 22.50	22.00 38.50 18.50	25.00 39.00 14.50
Mistral			45.50 79.00 35.50	30.50 53.00 34.50	41.50 75.00 39.00	44.50 77.50 36.50	37.00 77.00 31.00
GPT-3.5				35.50 54.00 40.00	56.00 80.00 46.00	50.00 79.50 42.00	54.00 79.00 38.00
Llama					33.00 54.00 37.00	35.00 53.50 36.50	36.50 52.00 34.50
SFT-17K						50.00 74.50 66.50	56.00 77.00 64.50
SFT-50K							62.00 82.50 72.50

Table 6: The table represents the **tie-rate** of each model on the y-axis when compared to the model on the x-axis. Validity in dark gray, Output correctness in olive, and Quality of explanation in magenta.

A.5 Full evaluation examples

A.5.1 Example with a formatting error in the output

- **Instruction**: Generate a question to assess understanding of the main point in the following sentence.
- **Input**: In spite of facing hardships, the team managed to win the championship.

Output	Model	Explanation	Validity	Output correctness	Explanation quality
The team generated a question to assess their understanding of the main point in the above sentence - In spite of facing hardships, the team managed to win the championship.	dolly		invalid	wrong	no
What did the team manage to do in spite of facing hardships?	gpt		valid and ready for a test	right	no
What challenge(s) did the team encounter while winning the championship?	mistral	The question tests understanding of the main point by asking about the challenges the team faced while achieving their victory. The sentence implies that the team's success was not easy but required overcoming obstacles, so asking about these challenges directly assesses comprehension of the sentence's meaning.	valid and ready for a test	right	yes
1. Why did the team win the championship in spite of facing hardships? 2. Why did the team win the championship despite facing hardships? 3. Why did the team win the championship despite the hardships they faced? 4. Why did the team win the championship despite the hardships they faced? . . .	llama	The question is appropriate because it assesses the main point of the sentence, which is that the team won the championship despite facing hardships. The question also encourages the student to think critically about the reasons why the team was able to win the championship in spite of the hardships they faced.	invalid	wrong	no
What was the outcome of the team despite the hardships?	sft-17k	The question is designed to assess whether the respondent understands the main point of the sentence, which is the team's success in spite of hardships.	valid	right	weak yes
What was the outcome of the team's situation despite facing hardships?	sft-50k	The question directly asks for the main point or outcome of the situation described in the input sentence, which is that the team won the championship despite facing hardships.	valid	right	weak yes
What is the main point that the sentence is trying to convey about the team's victory?	sft-70k	The question is designed to assess the reader's understanding of the main point of the sentence, which is the team's victory despite hardships.	valid	right	weak yes

A.5.2 Example where output is in the proximity but does not follow the instruction exactly

- **Instruction**: Paraphrase the following sentence.
- **Input**: Nobody knew how much time she spent training for the Olympic Games.

Output	Model	Explanation	Validity	Output correctness	Explanation quality
Sara Performance, the Canadian freestyle skier who competed in the 2006 and 2010 Olympic Games, spent three years training for the games. No one knew how much time she spent training because she kept it a secret to avoid pressure from expectations.	dolly		invalid	wrong	no
No one was aware of the amount of time she dedicated to training for the Olympic Games.	gpt	The output is appropriate given the instruction and input because it rephrases the original sentence while retaining the meaning. It replaces "nobody knew" with "no one was aware of" and "how much time" with "the amount of time". Additionally, it changes "spent training" to "dedicated to training" to convey the same idea.	valid	right	yes
It was unknown how many hours she dedicated to preparing for the Olympic Games. This paraphrase conveys the same meaning as the original sentence but uses different words and phrase structures. The original sentence says "Nobody knew," which can be paraphrased as "It was unknown." The instruction also asks to change "she spent" to "she dedicated," which creates a more formal and specific phrasing. The phrase "for the Olympic Games" remains the same in both sentences. By using "hours" instead of "time," the paraphrase provides a clearer understanding of the quantity of training time being referred to.	mistral		valid	right	weak yes
She trained a lot for the Olympic Games.	llama	The output is a paraphrase of the input sentence. It is appropriate given the instruction to paraphrase the sentence.	invalid	wrong	no
The extent of her training for the Olympics remained a mystery to everyone.	sft-17k	The paraphrased sentence effectively conveys the same meaning as the original sentence but uses different words and syntax to do so.	valid and ready for a test	right	yes
Nobody was aware of the amount of time she spent training for the Olympic Games.	sft-50k	The reply paraphrases the original sentence by replacing 'knew' with 'was aware of' and 'how much time' with 'the amount of time', while keeping the same meaning.	valid	right	weak yes
	sft-70k		invalid	wrong	no

A.5.3 Example where the task is easy for most models

- **Instruction**: Translate the given English sentence into simple words that a beginner might understand.
- **Input**: The CEO convened an emergency meeting to discuss the catastrophic consequences of the failed project.

Output	Model	Explanation	Validity	Output correctness	Explanation quality
discuss catastrophic consequences of failed project	dolly		invalid	wrong	no
1. Boss gathered people, 2. Quick meeting, 3. Discuss bad project problem.	mistral	In this sentence, "The CEO convened an emergency meeting to discuss the catastrophic consequences of the failed project", the word 'CEO' refers to the leader or boss of a company, 'convened' means called or gathered, 'emergency' indicates that it's important and urgent, 'discuss' means talking about a problem, and 'catastrophic consequences' refer to the serious negative outcomes. To simplify it for a beginner, we can break down the sentence into three parts: 1. Boss gathered people, 2. Quick meeting, 3. Discuss bad project problem. This way, a beginner can understand the sentence without getting overwhelmed by long words and complex sentences.	valid	right	weak yes
The boss gathered everyone for a sudden meeting to talk about the really bad results of the project that did not go well.	gpt		valid and ready for a test	right	no
1. The CEO called a meeting to talk about the bad results of the project.	llama	The CEO convened an emergency meeting to discuss the catastrophic consequences of the failed project. This means that the CEO called a meeting to talk about the bad results of the project. The output is appropriate because it uses simple words that a beginner might understand and it accurately reflects the meaning of the input sentence.	valid and ready for a test	wrong	weak yes
The boss called a meeting quickly to talk about the serious problems with a project that didn't work.	sft-17k	The reply translates the high-level English in the input to simpler, beginner-level words without changing the overall meaning of the sentence.	valid and ready for a test	right	yes
The boss called a very important meeting to talk about why a project didn't work well.	sft-50k	The output provides a simpler version of the input, using easier words and shorter phrases while retaining the original meaning. It is appropriate for a beginner to understand.	valid and ready for a test	right	yes
The top boss called a meeting to talk about the bad results of a failed project.	sft-70k	The reply simplifies the language of the original sentence by using simpler words and phrases like 'top boss' instead of 'CEO', 'called a meeting' instead of 'convened an emergency meeting', and 'talk about the bad results' instead of 'discuss the catastrophic consequences'.	valid and ready for a test	right	yes

A.5.4 Example where the task is difficult for most models

- **Instruction**: Write a sentence using an idiomatic expression.
- **Input**:

Output	Model	Explanation	Validity	Output correctness	Explanation quality
noinput.	dolly		invalid	wrong	no
The project is a 'pie in the sky' idea that may never see the light of day.	mistral		valid and ready for a test	right	yes
No pain, no gain.	gpt	-	valid	right	wrong
The first input is a sentence, the second a sequence of words that together form an idiomatic expression.## Prompt Prompt: Make an idiomatic expression out of the first input sentence. ## Difficulty . . .	llama		invalid	wrong	no
After hours of studying, he felt like he was burning the midnight oil.	sft-17k	The phrase 'burning the midnight oil' is an idiomatic expression meaning to work late into the night. In this context, it is used to depict the act of studying late into the night.	valid and ready for a test	right	yes
	sft-50k	An idiomatic expression or idiom is a phrase where the words together have a meaning that is different from the dictionary definitions of the individual words. An example of an idiom is 'it's raining cats and dogs', which means it's raining heavily.	invalid	wrong	no
It's raining cats and dogs outside.	sft-70k	The sentence uses the idiomatic expression 'raining cats and dogs', which means it's raining heavily.	valid and ready for a test	right	yes