Fine-tuning with `HED-IT`:
The impact of human post-editing for dialogical language models

Daniela Occhipinti[1,2], **Michele Marchi**[1,2], **Irene Mondella**[3,4], **Huiyuan Lai**[3],
Felice Dell'Orletta[4], **Malvina Nissim**[3], **Marco Guerini**[1]

[1]Fondazione Bruno Kessler, Italy, [2]University of Trento, Italy
[3]University of Groningen, Netherlands, [4]ItaliaNLP Lab @ CNR-ILC, Italy
{docchipinti, mmarchi, guerini}@fbk.eu, i.mondella@studenti.unipi.it.
felice.dellorletta@ilc.cnr.it, {h.lai, m.nissim}@rug.nl

Abstract

Automatic methods for generating and gathering linguistic data have proven effective for fine-tuning Language Models (LMs) in languages less resourced than English. Still, while there has been emphasis on data quantity, less attention has been given to its quality. In this work, we investigate the impact of human intervention on machine-generated data when fine-tuning dialogical models. In particular, we study (1) whether post-edited dialogues exhibit higher perceived quality compared to the originals that were automatically generated; (2) whether fine-tuning with post-edited dialogues results in noticeable differences in the generated outputs; and (3) whether post-edited dialogues influence the outcomes when considering the parameter size of the LMs. To this end we created `HED-IT`, a large-scale dataset where machine-generated dialogues are paired with the version post-edited by humans. Using both the edited and unedited portions of `HED-IT`, we fine-tuned three different sizes of an LM. Results from both human and automatic evaluation show that the different quality of training data is clearly perceived and it has an impact also on the models trained on such data. Additionally, our findings indicate that larger models are less sensitive to data quality, whereas this has a crucial impact on smaller models. These results enhance our comprehension of the impact of human intervention on training data in the development of high-quality LMs.

1 Introduction

The landscape of language technology is constantly being populated by new instruction-based and conversational models, thanks to the increased availability of foundational Large Language Models (LLMs) and easy access to fine-tuning support. Especially for lesser-resourced languages, the standard procedure is to automatically translate existing English datasets and use those to fine-tune one
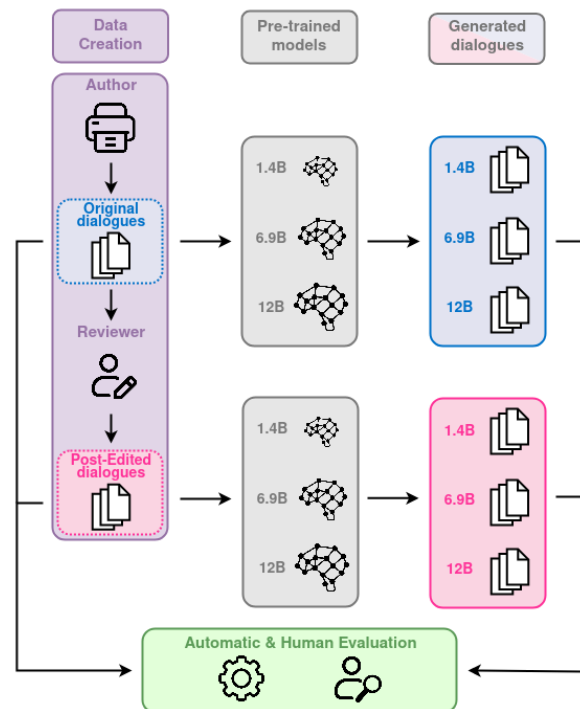

Figure 1: Data creation pipeline. The *author* (machine) provides initial dialogues, the *reviewer* (human) post-edits them. These two datasets are then used to train different sizes of an LM. Automatic and human evaluation is performed both on training data and corresponding trained models to assess the impact of curation.

of the available LMs on the target language (see Naveed et al. (2023) for a recent overview of models and procedures).

Still, this proliferation of models is not matched by progress in evaluation efforts aimed at understanding exact differences, especially in terms of data and model quality as perceived by humans (rather than just overall model performance). Gaining such insights would help in optimising the balance between cheaper and more costly choices, such as pure automatic translation vs manual intervention, in the creation of datasets for developing high-quality models. This is the aim of the present

contribution, focusing on Italian dialogic models, but applicable to other under-resourced languages. By building on previous research revolving around the concept of data quality, we adopted an *author-reviewer* architecture (Tekiroğlu et al., 2020) to create the `HED-IT` (Human Edited Dialogues for ITalian) dataset[1] where automatically generated dialogues are post-edited by humans. We then use such edited dialogues, as well as the unedited, to fine-tune an LLM in three different sizes. This setup enables us to study the impact of human intervention on training data (vs its absence) on the final resulting models. We do so in terms of (i) quality assessment by humans of LLM output; and (ii) behaviour assessment across model sizes. Figure 1 provides an overview of our approach. The research questions posed and addressed in this paper can be summarised as follows:

Q1 We investigate whether post-edited dialogues to be used for fine-tuning are (perceived as) better in quality than the original (automatically generated) ones. Automatic metrics and dedicated manual evaluation suggest that this is the case.

Q2 We investigate if fine-tuning with post-edited data yield detectable differences in the fine-tuned models. Model evaluation via automatic metrics and output evaluation via human judgement indicate an overall quality improvement for the models fine-tuned with post-edited data.

Q3 We study the impact of the LLM's size and observe that the impact of post-edited data is stronger for smaller model sizes.

In addition to providing a dataset of more than 16,000 original and post-edited Italian dialogues, and a battery of fine-tuned models of different sizes, we contribute insights and a methodology that can be applied to other languages and tasks.

2   Related Work

LLMs for Conversational Models   LLMs with billions of parameters, pre-trained on a huge amounts of language data in a self-supervised manner, achieved great success (Brown et al., 2020; Naveed et al., 2023), especially as the backbone of conversational agents often achieving human-like interaction (Thoppilan et al., 2022; OpenAI, 2022; Wei et al., 2023). However, no open-source

model, especially in languages other than English, comes close in capabilities to closed-source models, and creating instruction-following data for fine-tuning can be highly time-consuming and costly. This has prompted research into finding ways to reduce such efforts. For instance, Wang et al. (2023) proposed a framework to fine-tune LLMs using instructional signals from the model itself, reducing the need of human-written instructions. Furthermore, knowledge distillation (Hinton et al., 2015), which involves transferring knowledge from larger to smaller models, has been used to train open-source conversational agents (Xu et al., 2023; Chiang et al., 2023). Still, instruction-based or conversational models in languages other than English are typically trained using automatically translated data (Chen et al., 2023b; Larcher et al., 2023; Ranaldi et al., 2023). This data, mostly in the form of instructions, originates in English. For instance, a suite of LLaMA-based models in various languages has been trained using the same dataset that was initially developed and utilised for English Alpaca (Taori et al., 2023). In this work, we specifically focus on the Italian landscape.

Italian Conversational Models   The Italian model corresponding to English Alpaca (Taori et al., 2023) is Camoscio (Santilli and Rodolà, 2023), which is trained with low-rank adaptation (LoRA, Hu et al. 2022) on the Alpaca dataset automatically translated to Italian. Stambecco[2] is an Italian instruction-following model, based on LLaMA (Touvron et al., 2023a) and fine-tuned on the Italian translation of GPT-4-LLM dataset (Peng et al., 2023). Fauno (Bacciu et al., 2023) is a more conversation-like Italian model, trained on English data translated to Italian, originally generated by ChatGPT chatting with itself. Most recently, Basile et al. (2023) released LLaMAntino-Chat, a family of models also based on the LLaMA 2 Chat (Touvron et al., 2023b), adapted on the Italian language by fine-tuning on the Italian translation of the UltraChat dataset (Ding et al., 2023).

Low-Resource Languages Dialogue Datasets Several approaches have been adopted in literature to collect multilingual or low-resource language datasets. Various methodologies automatically translate English data to the language of interest (Bacciu et al., 2023; Santilli and Rodolà,

---

2023; Chen et al., 2023b; Larcher et al., 2023; Ranaldi et al., 2023). In their work, Hu et al. (2023) propose the Multi³WOz dataset, a comprehensive multilingual, multi-domain, task-oriented dataset with dialogues in four languages: English, Arabic, French, and Turkish. These dialogues were obtained using a Wizard of Oz (WoZ) methodology, where bilingual speakers, native in the target language and fluent in English, crafted natural and culturally relevant conversations. However, this method was resource-intensive, consuming significant human and financial resources and taking over a year to compile the dialogues. Another dialogic dataset, both in English and Chinese, is BiToD (Lin et al., 2021). BiToD is a task-oriented dataset in which dialogues are collected through a dialogue simulator interacting with a knowledge base to generate dialogue outlines that are then revised by crowdsourcers to convert them into natural conversations. While this method resembles our proposed approach, the goals differ significantly. Lin et al. (2021) are interested in the bilingual and cross-lingual evaluation of the task. In contrast, we focus on in-depth analyses of a low-resource language and open-domain dataset with original and revised versions, assessing data quality, collection efficiency, and their effects on model fine-tuning across different sizes.

Data Quality The quality of fine-tuning plays a crucial role in improving performance with less data (Zhou et al., 2023a). While the benefits of post-editing in machine translation are long-known (Castilho et al., 2014; Castilho and O'Brien, 2016), in developing LLMs the main focus has been on data curation (document filtering) to increase quality (Zhou et al., 2023b; Li et al., 2023; Wang et al., 2024; Furman et al., 2023), while less attention has been given to document correction via post-editing (Tekiroğlu et al., 2022).

3 Dataset Creation

To create the HED-IT dataset, we used the *author-reviewer* pipeline proposed by Tekiroğlu et al. (2020), in which an LLM (the *author* component) produces novel data, while humans (the *reviewer* component) filter and eventually post-edit them. We tested several author configurations and introduced two adaptations to the original proposal: (i) following Russo et al. (2023), rather than fine-tuning a GPT model, we opted for an instruction-based LLM (i.e., ChatGPT (OpenAI, 2022)) that

does not require fine-tuning; (ii) we extend the concept of a machine-based *author* in this setting to include also the automatic extraction of excerpts from existing human-produced text. To create the dataset, we used non-real data sourced from movie scripts and synthetic data to address potential privacy and ephemerality concerns associated with real data (Klubicka and Fernández, 2018).

The following sections provide a detailed overview of the sources and methodologies used to construct the HED-IT dataset.

3.1 Author Module

For the author module, we tested three strategies, whose common feature is that they do not rely on the manual creation of new dialogues from scratch, which is particularly costly: (i) automatically extracting excerpts from human-written material, (ii) using an LLM to rewrite human-written dialogues, (iii) automatically generating dialogues prompting an LLM with human-written material. Our aim in using these configurations was to create a dataset with a wide range of linguistic characteristics and topics. We constrained the number of speakers to two and the number of turns to a minimum of three to maintain clarity and consistency, avoiding complex multi-character conversations.

Human Dialogues (Dial$_H$) To obtain human-written dialogues without having to produce them anew, we exploited movie scripts scraped from various sources, with no genre-specific filters. In this case, the author module is an algorithm that extracts script portions meeting some given characteristics (e.g. excerpts between two speakers, minimum three turns, further details are in Appendix A.1).

Human-LLM Dialogues (Dial$_{H+LLM}$) One limitation of extracting portions of movie dialogues is that, by nature, they rely heavily on visual support, which can make them less understandable when taken out of context. To reintroduce some contextual depth to the extracted dialogues, and to make them self-contained, we instructed Chat-GPT to rewrite them in a more comprehensible way (details and prompt in Appendix A.2).

LLM Dialogues (Dial$_{LLM}$) Gathering new Italian dialogues can be costly, and utilizing real data poses potential privacy issues. Thus, following the approach suggested by Chen et al. (2023a), we generated synthetic dialogues by instructing Chat-GPT with diverse prompts using various contexts

(further details in Appendix A.3). The contexts provided in the prompt were obtained from four distinct sources: (i) a set of background information about two speakers and a topic sourced from PLACES (Chen et al., 2023a), automatically translated from English to Italian using DeepL; (ii) a set of conversation-starting questions obtained from a website which supports social interaction; (iii) a set of tweets from the X/Twitter account of ANSA[3], one of the main news agencies in Italy; (iv) a subset of the Twitter Italian Dialect Data dataset[4].

3.2 Reviewer Module

Two Italian native speakers (MSc students in Computational Linguistics) were hired as annotators for the post-editing process. Following the methodology proposed by Fanton et al. (2021), both annotators underwent extensive training on the data and task, including initial sessions to establish guidelines. In these sessions, they were given a selection of author-generated dialogues to identify common issues and to propose the most effective strategies for enhancing their quality. Subsequently, these strategies were integrated into the guidelines. Once the annotation campaign started, weekly meetings were held to address any possible issues or feedback from annotators.

Post-editing guidelines aimed to reduce annotators' effort while maintaining the quality of the output. Below, we discuss examples of common phenomena and post-editing actions, while in Table 1, we present an example of dialogue post-editing.

Post-Editing of Dial$_H$ Excerpts extracted from movie scripts present unique phenomena due to their specific linguistic characteristics, including dialect usage, visual references, and fictional situations common in genres like horror, fantasy, and sci-fi. Therefore, tailored guidelines are necessary:

- Overly generic or difficult-to-understand dialogues, such as *Beh, e... / Se ne hai voglia, ne posso trovare quanta ne vuoi. / Come?* (EN: *Well, and... / If you feel like it, I can find as much as you want. / How?*), requiring excessive effort for post-editing, should be removed.

- Sometimes, dialogues may contain dialectal terms. If the terms are comprehensible to those unfamiliar with the dialect used, they can be kept,

Original	EN
A: ~~Mettetevi in circolo.~~ Lei chi è?	~~Make a circle.~~ Who are you?
B: Sono il marito.	I am the husband.
A: Spiacente, soltanto le persone ~~interessate~~ sono ammesse qui.	I am sorry, only ~~interested~~ people are allowed here.
Post-Edited	**EN**
A: Lei chi è?	Who are you?
B: Sono il marito.	I am the husband.
A: Spiacente, soltanto le persone coinvolte nel caso sono ammesse qui.	I am sorry, only people involved in the case are allowed here.
B: Ma io sono suo marito! Non mi potete negare l'accesso.	But I am her husband! You cannot negate me the access.

Table 1: Example of dialogue post-editing. In red the deletions and in green the additions.

such as the Tuscan *O questo chi l'è?* (EN: *Who is this?*). Otherwise, they should be removed.

- Interjections typical of the spoken language (e.g. *Ah!*, *Mmm...*, *Eh!*) can be kept.

- To maintain coherence, overly long dialogues or including abrupt topic changes should be split.

- Dialogues may have unclear or unhelpful turns, which should be removed for clarity. If a turn is deleted, either the preceding or following turn must be removed or merged to maintain the alternating speakers.

- Sometimes, dialogues may lack initial context or a proper ending. In such instances, annotators can create and insert a beginning and/or concluding turn to ensure coherence in the final dialogue.

Post-Editing of Dial$_{H+LLM}$ and Dial$_{LLM}$ Specific guidelines were necessary for post-editing dialogues either rewritten or fully generated by ChatGPT. Key issues include repetitive and redundant content, as well as excessive use of politeness characterized by frequent agreement and continuous expressions of gratitude. This tendency towards politeness might be attributed to the use of Reinforcement Learning With Human Feedback (Stiennon et al., 2020), which encourages extremely cautious and gentle responses, often resulting in unrealistic dialogues. Even when instructed to express strong disagreement, the model tends to produce excessive level of politeness.

- Repetitive phrases such as *"Capisco il tuo punto di vista, ma..."* (EN: *"I understand your point of*

---

view, but...") and *"Non capisco come tu possa..."* (EN: *I don't understand how you can...*) should be either rephrased or deleted.

- Generated dialogues can include redundant turns or information, e.g., *Su questo siamo d'accordo. Grazie per la tua opinione, Marco. / Di nulla, Alice. È sempre interessante discutere di questi temi con te.* (EN: *We agree on this. Thank you for your opinion, Marco. / You're welcome, Alice. It is always interesting to discuss these topics with you.*). This content should be deleted.

- Generated dialogues sometimes appear excessively polite, with frequent agreement and continuous expressions of gratitude, such as *Sì, hai ragione. Ma.. / Sì, forse hai ragione. Ma... / Assolutamente d'accordo. Ma...* (EN: *Yes, you are right. But... / Yes, you may be right. But... / Absolutely agree. But...*) Such instances should be curbed as much as possible.

- The endings of the generated dialogues often appear repetitive and standardised. These closures should be reworded as much as possible.

4 Dataset Analysis

Following the author-reviewer process, two versions of our dataset were collected: one with the original dialogues (HED-IT$_{orig}$) and another with the post-edited versions (HED-IT$_{p\text{-}e}$). HED-IT$_{orig}$ comprises over 9000 dialogues and more than 66000 turns, whereas HED-IT$_{p\text{-}e}$ comprises over 7000 dialogues and almost 48000 turns. On average, dialogues contain 8 turns. More details are reported in Table 2.

	Dial$_H$	Dial$_{H+LLM}$	Dial$_{LLM}$	Total
Dial$_{orig}$	2504	3458	3339	9301
Dial$_{p\text{-}e}$	1107	2824	3266	7197
Turns$_{orig}$	14042	22214	30034	66290
Turns$_{p\text{-}e}$	7817	16422	23704	47943
Turns/Dial$_{orig}$	6	7	12	8
Turns/Dial$_{p\text{-}e}$	8	7	10	8
Tok/Dial$_{orig}$	138	164	424	251
Tok/Dial$_{p\text{-}e}$	173	140	332	232
Tok/Turn$_{orig}$	24	24	37	31
Tok/Turn$_{p\text{-}e}$	23	22	35	29

Table 2: Dimensions of the HED-IT$_{orig}$ and HED-IT$_{p\text{-}e}$ datasets in terms of dialogues and turns. The upper part reports counts, while the lower reports averages.

To assess the validity of the methodology employed to build HED-IT, we also conducted an ad-

ditional analysis where a human annotator created a set of dialogues from scratch. We then compared the time needed for post-editing a set of dialogues versus composing them from the beginning. The results are summarised in Table 3.

	From scratch	Post-editing
Dial. per hour	17	**39**
Turns per hour	158	**381**
Tok. per hour	1695	**6627**

Table 3: Time needed (in seconds) for dialogues written from scratch by a single human annotator vs. machine-generated dialogues post-edited by the same annotator.

The table demonstrates a marked productivity gain between dialogues written by a human annotator and those that are machine-generated and then post-edited by the same annotator. The latter results in more than double the dialogues and turns processed, and nearly four times the tokens, highlighting a significant efficiency improvement, crucial for dataset scalability[5]. This is particularly substantial if we compare our proposed methodology to the WoZ scenario as in Multi³WOz, where not only dialogues are written from scratch, but also each simulated turn causes one participant to wait in an "idle state" for the other to answer, theoretically doubling the annotator's required time.

To assess the impact of post-editing both quantitatively and qualitatively, we use automatic metrics and a human evaluation.

4.1 Automatic Metrics

Inspired by Russo et al. (2023), we evaluated the quality of the post-edited data compared to the original data employing the *Human-targeted Translation Edit Rate* (HTER; Snover et al. (2006)) and the *Repetition Rate* (RR; Bertoldi et al. (2013)).

HTER measures the number of editing required to transform a machine-generated text into its post-edited version. The score is computed considering word substitution, insertions and deletions.

RR is used to assess how repetitive a text is. It calculates the geometric mean of the frequency of repeated sequences of words (*n-grams*) within the text. To ensure consistency across texts of varying lengths, a fixed-size sliding window is used during processing. In our analysis, we focused on word *n-grams*, where *n* varies from 1 to 4, and employed

---

[5]A similar improvement using an *author-reviewer* pipeline is reported in (Russo et al., 2023).

a sliding window of 1000 words.

Table 4 summarises the impact of human intervention across the different portions of the dataset. Several phenomena can be seen. (i) Dial_H have a higher deletion rate than other authors' output, both in terms of whole dialogues (0.559 vs 0.191 and 0.022) and turns (0.436 vs 0.249 and 0.186). A post-hoc interview with the annotator showed that this is mainly due to their being "decontextualized excerpts" and over-reliant on visual context, making them unsuitable for simple post-editing. (ii) In contrast, $\text{Dial}_\text{H+LLM}$ and Dial_LLM tend to have fewer deletions. This implies that more dialogues and turns are post-edited (0.373 vs 0.761 and 0.954). Similarly, the percentage of post-edited turns is higher. Interestingly, the required post-editing is typically less intensive (HTER of 0.616 vs HTER of 0.358). (iii) However, we also observed that Dial_LLM have a more "artificial" nature: there is an increase in RR as compared to other author configurations (4.771 vs 2.603 and 2.467). (iv) Finally, there is a clear beneficial effect of the post-editing, that consistently enhances the quality of dialogues by reducing the RR in all configurations.

	Dial_H	$\text{Dial}_\text{H+LLM}$	Dial_LLM	Total
Dial_{unch}	**0.068**	0.048	0.024	0.045
Dial_{del}	0.559	0.191	**0.022**	0.229
Dial_{edit}	0.373	0.761	**0.954**	0.726
Turns_{unch}	0.475	0.480	**0.524**	0.501
Turns_{del}	0.436	0.249	**0.186**	0.253
Turns_{edit}	0.089	0.271	**0.290**	0.253
HTER_{edit}	0.616	0.611	**0.358**	0.459
RR_{orig}	2.603	**2.467**	4.771	3.790
$\text{RR}_{p\text{-}e}$	2.428	**1.992**	3.590	2.965

Table 4: Analysis of Post-Editing interventions. *unch*: turn/dialogue was not changed; *del* turn/dialogues was removed during post-editing; *edit*: intervention that was not deletion. HTER is computed only on edited turns.

4.2 Human Evaluation

Besides the automatic metrics, we used human evaluation to assess the impact of post-editing. We selected a total of 416 dialogues from the Dial_LLM subset: 208 post-edited dialogues (half by one annotator and half by the other) and their corresponding original versions. Dialogues from movie scripts and the version rewritten by the author model were excluded as they may be easily recognisable, leading to possible bias into the responses.

The sample was evaluated by native Italian speakers recruited via Prolific[6], a platform specifically designed for research purposes. Each evaluator received a survey consisting of 8 dialogues stratified by length: *short dialogues* with 10 or fewer turns, *medium* with between 10 and 15 turns and *long* with more than 15 turns. To cover the 416 dialogues, we created 52 surveys, each assigned to 3 different evaluators. In total, we had almost 2500 evaluations from 129 different participants[7].

The surveys aimed to assess the dialogues along two dimensions: understandability and naturalness (defined as the "likelihood that it was written by a machine"), both using a 5-point scale. The responses were categorised into evaluations regarding the comprehensibility and naturalness of both the Dial_{orig} and $\text{Dial}_{p\text{-}e}$ dialogues. The mean response was then calculated for each subgroup (micro-average). Results are reported in Table 5: the mean Understandability of original dialogues is lower than for post-edited dialogues (4.045 and 4.175 respectively), while Machine Probability is higher (3.296 vs 2.875). This shows that post-editing leads to more understandable and less artificial dialogues. Furthermore, comparing each original dialogue directly with its post-edited version and calculating the average responses of every dialogue, shows 0.130 increase in understandability and a 0.421 decrease in artificiality, on average.

		Understand ↑	Mach. Prob. ↓
orig	Low edit	4.122	3.160
	High edit	3.968	3.433
	Overall	4.045	3.296
p-e	Low edit	4.266	2.913
	High edit	4.083	2.837
	Overall	**4.175**	**2.875**

Table 5: Human evaluation answers' micro-averages.

Dialogues were also sorted by post-editing rate, dividing them into *highly post-edited* and *lowly post-edited*. To do so, we took into account the HTER and the number of eliminated turns for each dialogue. Table 5 shows also that, for the second question (Machine Probability), the difference between original and post-edited is greater for highly post-edited dialogues than for lowly post-edited ones (a reduction of 0.596 vs a reduction of 0.247).

This is caused by the fact that the highly post-edited dialogues were perceived as very unnatural in their original version, and therefore benefited greatly from post-editing (the more post-editing, the greater the improvement in naturalness).

5 Experimental Design

Using the HED-IT dataset, our objective is to understand the impact of post-editing on the performances of an LLM. We evaluated the models automatically using the test set gold standard, and ran human evaluations of models' outputs.

5.1 Dataset

To obtain two datasets of equal sizes, we started from the HED-IT$_{p-e}$, which has fewer dialogues compared to HED-IT$_{orig}$, and split it into train, validation and test subsets (80:10:10), stratified according to author-strategy (i.e., Dial$_H$, Dial$_{H+LLM}$, and Dial$_{LLM}$). We then subsampled the same number of dialogues from the HED-IT$_{orig}$ partition including original dialogues that had been deleted during post-editing. In this way, we both controlled for the training data size and granted that the distribution of subsampled data represents the distribution of the real *Original* data, i.e. including low-quality dialogues as well. After the subsampling, 83% of the original dialogues were matched with their post-edited counterpart, while the remaining 17% came from the "deleted" set.

5.2 Models

As base models, we use Pythia (Biderman et al., 2023), a suite of decoder-only autoregressive language models. Pythia's LLMs are trained on publicly available data, maintaining the same order of training data and varying only in size from 70M to 12B. This setup allows us to isolate the impact of post-editing on the LLMs' generations, considering the LLMs dimensions as the only variable[8].

For this study, we focused on the Pythia sizes of 1.4B, 6.9B and 12B. We conducted a preliminary experiment to assess the proficiency of these LLMs in Italian in a zero-shot setting, which showed that these models are suitable for our experiments. Subsequently, we fine-tuned these models with QLoRA (Dettmers et al., 2023) in six different configurations. These configurations involved each of the

---

[8]While other LMs are available for Italian that could be used for fine-tuning, Pythia is the only one that allows us to run controlled experiments including size, and rule out the effect of other variables.

three Pythia model sizes, fine-tuned on either the *Original* and the *Post-Edited* dialogues (see Figure 1). Furthermore, we ensured that the models were fine-tuned with the same data order, consistently with Pythia's methodology. For details on fine-tuning configurations, see Appendix C.1.

5.3 Model Evaluation

To assess the quality of the models we used the following automatic metrics:

- **Conditional turn Perplexity (CPPL)** (Su et al., 2021; Occhipinti et al., 2023) measures the perplexity of a gold turn given the previous dialogue history. We computed the CPPL for each turn iteratively and considered the average.

- **Average Accuracy at N (Acc@N)** Welch et al. (2022) defines the prediction of a word from a gold turn being correct if it appears within the top N most probable words provided by the model. We computed the Acc@N at each turn incrementally, considering the average across all turns.

Models fine-tuned on original data were tested on original data, while those fine-tuned on post-edited data were tested on post-edited data.

5.4 Output Generation

To let the models generate dialogues we prompt them with the first two turns of a subset of 100 dialogues which have been used for the human evaluation of the dataset (Section 4.2). For generation, we used the *top-p* decoding mechanism with a value of 0.9, a temperature of 1, and a repetition penalty of 2. As we use six models, we obtain a total of 600 generated dialogues.

We noticed a consistent pattern in all model-generated dialogues: since we forced the model to generate content extensively, eventually, it reached a limit where it began repeating turns or using synonyms excessively. This resulted in unnatural and inconsistent dialogue. We call this phenomenon "derailment". To automatically detect the start of derailment in each model's output, we used BLEU (Papineni et al., 2002) with a threshold of 0.9 to identify when the turns become overly similar. Table 6 shows the average length of dialogues and turns when the derailment is cut off. Models trained on post-edited dialogues appear to derail sooner compared to models trained on original dialogues. This could be attributed to the length discrepancy between post-edited and original dialogues: during

fine-tuning, models trained on post-edited data are exposed to shorter dialogues (251 vs 232 token on average per dialogue, see Table 2). Consequently, at inference time, they might find it challenging to generate dialogues as lengthy as those produced by models trained on original dialogues.

Model	Turns	Tok/Dial	Tok/Turn
$1.4B_{orig}$	14	267	19
$1.4B_{p-e}$	13	251	19
$6.9B_{orig}$	13	394	31
$6.9B_{p-e}$	12	314	27
$12B_{orig}$	21	420	20
$12B_{p-e}$	14	402	29

Table 6: Average lengths of generated dialogues considering BLEU score 0.9 derailment cut.

5.5 Output Evaluation

Human evaluation was carried out in the same manner as for the dataset analysis, namely with crowdsourced assessments over understandability and probability of automatic generation (Section 4.2). Derailment turns were excluded, and dialogues were cut at 20 turns to prevent negative impacts from derailment and excessive length.

6 Results

In this section, we will outline the results obtained on the six configurations we tested.

6.1 Automatic Evaluation

The CPPL scores in Table 7 show that models trained on post-edited dialogues consistently outperform those trained on original dialogues across all Pythia dimensions. Notably, the CPPL difference between models trained on original and post-edited dialogues is lower for the biggest model (a difference of 2.462 for 1.4B vs a difference of 0.748 for 12B), indicating a more significant impact of post-editing on smaller models.

Model	CCPL	Δ
$1.4B_{orig}$	14.016	2.462
$1.4B_{p-e}$	11.554	
$6.9B_{orig}$	13.098	2.663
$6.9B_{p-e}$	**10.465**	
$12B_{orig}$	11.733	0.748
$12B_{p-e}$	10.985	

Table 7: Models' CPPL values over gold turns.

Table 8 presents Acc@N scores for various configurations. Acc@N values are higher for models trained on original dialogues compared to post-edited dialogues across all turns (e.g., 0.627 of Acc@1 for $12B_{orig}$ vs 0.612 of Acc@1 for $12B_{p-e}$). We hypothesise that the presence of excessively polite responses generated by the Author model (ChatGPT) in original dialogues, often edited out or rephrased in post-edited versions, may contribute to this difference, especially considering the prevalence of $Dial_{H+LLM}$ and $Dial_{LLM}$ dialogues. To test this hypothesis, we recomputed the metrics systematically excluding the last 20% and 30% of each test set dialogue. We indeed observed a decrease in Acc@N for original dialogues, leading to a convergence in performance between original and post-edited dialogues (e.g. 0.621 of Acc@1 for $12B_{orig}$ vs 0.619 of Acc@1 for $12B_{p-e}$).

The output of the models has been automatically assessed using the RR metric. The results are presented in Table 9. The scores demonstrate that removing the derailment consistently improves RR scores in all configurations. Moreover, results show that the impact of the post-editing is more visible in smaller models. In fact, smaller models trained on post-edited dialogues have lower RR rates compared to those trained on original dialogues (e.g., 1.434 for the 1.4B model trained on post-edited dialogues vs 1.534 for the same model trained on original dialogues). Conversely, for the largest models, the RR is higher for the model trained on original dialogues compared to its counterpart trained on post-edited dialogues (i.e., 1.490 vs 1.781).

6.2 Human Evaluation

For the human evaluation of the generated dialogues, we selected 600 dialogues (100 for each model), resulting in 75 different surveys submitted for crowdsourced ratings. Each dialogue is assessed by three participants. Participants from the first human evaluation (Section 4.2) were excluded from this round. Examples of well and poorly rated dialogues are given in Appendix D. In total, we had 3600 evaluations from 169 different participants.

Results are in Table 10. Models trained on post-edited dialogues generate text perceived as more natural, as the Machine Probability score for these models is always lower than for models trained on the original dialogues (4.137 of $1.4b_{p-e}$ vs 4.237 of $1.4b_{orig}$, 3.887 vs 4.037 for the 6.9b model and 3.927 vs 4.077 for the 12b one). Understandability improves with post-editing only for the largest

Model	All turns			-20%			-30%		
	Acc@10	Acc@5	Acc@1	Acc@10	Acc@5	Acc@1	Acc@10	Acc@5	Acc@1
1.4B$_{orig}$	0.823	0.769	0.583	0.823	0.767	0.582	0.818	0.763	0.578
1.4B$_{p\text{-}e}$	0.817	0.761	0.571	0.821	0.765	0.578	0.818	0.763	0.577
6.9B$_{orig}$	0.855	0.805	0.618	0.854	0.804	0.616	0.850	0.799	0.612
6.9B$_{p\text{-}e}$	0.850	0.798	0.608	0.855	0.803	0.616	0.852	0.800	0.615
12B$_{orig}$	**0.863**	**0.815**	**0.627**	**0.862**	**0.811**	**0.625**	**0.858**	**0.807**	**0.621**
12B$_{p\text{-}e}$	0.852	0.802	0.612	0.857	0.806	0.620	0.855	0.804	0.619

Table 8: Models' Acc@N over gold turns, with 20% and 30% of dialogues' last turns progressively removed.

	Original		Post-Edited	
	pre-cut	post-cut	pre-cut	post-cut
1.4B	1.888	1.534	1.651	**1.434**
6.9B	2.637	2.022	2.149	**1.705**
12B	1.928	**1.490**	2.202	1.781

Table 9: Repetition Rate (RR) over generated dialogues.

model among those considered (3.493 of 12b$_{p\text{-}e}$ vs 3.447 of 12b$_{orig}$), while the other two models do not seem to improve when trained on post-edited dialogues. In summary, humans perceive the impact of post-editing training data on model generation.

	Understand ↑	Machine Prob. ↓
1.4b$_{orig}$	2.987	4.237
1.4b$_{p\text{-}e}$	2.977	4.137
6.9b$_{orig}$	3.430	4.037
6.9b$_{p\text{-}e}$	3.377	**3.887**
12b$_{orig}$	3.447	4.077
12b$_{p\text{-}e}$	**3.493**	3.927

Table 10: Human evaluation of generated dialogues.

7 Conclusions

Machine-generated data is commonly used to fine-tune LLMs in low-resourced languages, while improving data quality has received little attention. To this end, we extensively investigated the impact of human intervention on data used for fine-tuning dialogical models. We provided a corpus of machine-generated Italian dialogues paired with post-edited counterparts, and trained an LLM in three different sizes using this corpus. Results clearly showed that post-edited dialogues had a higher perceived quality compared to their original counterparts. Model evaluation also indicates an overall quality improvement for the models fine-tuned with post-edited data. Still, results suggest that the effect of post-editing is stronger for smaller models while slowly

fading for bigger ones, and this can be a critical aspect when running a data collection campaign. Also the intervention of machine-rewriting before human post-editing can be beneficial when available training data, even if human-written, is silver with respect to the target domain.

Limitations

This work has a few limitations worth noting. Firstly, the resource presented is available only in Italian. However, it is important to highlight that the author-reviewer approach used for data collection is language-independent and can be replicated assuming that an LLM for the target language is available.

A significant portion of the data was generated by prompting an LLM (ChatGPT): it is important to recognise that prompting autoregressive models can bring to uncontrolled forms of generation, such as hallucinations that for some scenarios might be critical.

Additionally, it is important to acknowledge that our testing was limited to only a subset of possible models (i.e. Pythia), although we would not expect great differences in results with other models.

Finally, while human post-editing can greatly improve data quality, it might require substantial resources in terms of human effort, time, and financial investment if author module's outputs are not good enough.

Ethics Statement

In this study, two of the authors acted as annotators in the post-editing process. For human evaluation, we used crowdsourcing via the Prolific platform. All crowdsourced workers were compensated following the platform's ethical payment principles[9].

---

[9] https://researcher-help.
prolific.com/hc/en-gb/articles/
4407695146002-Prolific-s-payment-principles

Finally, it is essential to highlight that the study did not assess the safety of the models used, but our fine-tuning data were all validated by the annotators, so any safety concern is mainly due to pre-training data.

References

Andrea Bacciu, Giovanni Trappolini, Andrea Santilli, Emanuele Rodolà, and Fabrizio Silvestri. 2023. Fauno: The italian large language model that will leave you senza parole! *arXiv preprint arXiv:2306.14457*.

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language. *arXiv preprint arXiv:2312.09993*.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *Proceedings of Machine Translation Summit XIV: Papers*, Nice, France.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sheila Castilho and Sharon O'Brien. 2016. Evaluating the impact of light post-editing on usability. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 310–316, Portorož, Slovenia. European Language Resources Association (ELRA).

Sheila Castilho, Sharon O'Brien, Fabio Alves, and Morgan O'Brien. 2014. Does post-editing increase usability? a study with Brazilian Portuguese as target language. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 183–190, Dubrovnik, Croatia. European Association for Machine Translation.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023a. PLACES: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Barry Haddow, and Kenneth Heafield. 2023b. Monolingual or multilingual instruction tuning: Which makes a better alpaca. *arXiv preprint arXiv:2309.08958*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

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*.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Damián Furman, Pablo Torres, José Rodríguez, Diego Letzen, Maria Martinez, and Laura Alemany. 2023. High-quality argumentative information in low resources approaches improve counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2942–2956, Singapore. Association for Computational Linguistics.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023. Multi3WOZ: A Multilingual, Multi-Domain, Multi-Parallel Dataset for Training and Evaluating Culturally Adapted Task-Oriented Dialog Systems. *Transactions of the Association for Computational Linguistics*, 11:1396–1415.

Filip Klubicka and Raquel Fernández. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. In *4REAL 2018 Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language*, page 16.

Celio Larcher, Marcos Piau, Paulo Finardi, Pedro Gengo, Piero Esposito, and Vinicius Caridá. 2023. Cabrita: closing the gap for foreign languages. *arXiv preprint arXiv:2308.11878*.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. *arXiv preprint arXiv:2106.02787*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Daniela Occhipinti, Serra Sinem Tekiroglu, and Marco Guerini. 2023. Prodigy: a profile-based dialogue generation dataset.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI blog*.

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.

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

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*.

Daniel Russo, Shane Kaszefski-Yaschuk, Jacopo Staiano, and Marco Guerini. 2023. Countering misinformation via emotional response generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11476–11492, Singapore. Association for Computational Linguistics.

Andrea Santilli and Emanuele Rodolà. 2023. Camoscio: an italian instruction-tuned llama.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

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

Hsuan Su, Jiun-Hao Jhan, Fan-yun Sun, Saurav Sahay, and Hung-yi Lee. 2021. Put chatbot into its interlocutor's shoes: New framework to learn chatbot responding with intention. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1559–1569, Online. Association for Computational Linguistics.

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.

Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114, Dublin, Ireland. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 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*.

Jiahao Wang, Bolin Zhang, Qianlong Du, Jiajun Zhang, and Dianhui Chu. 2024. A survey on data selection for llm instruction tuning.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. 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, Toronto, Canada. Association for Computational Linguistics.

Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2023. Leveraging large language models to power chatbots for collecting user self-reported data.

Charles Welch, Chenxi Gu, Jonathan K. Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. 2022. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, Dublin, Ireland. Association for Computational Linguistics.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6268–6278, Singapore. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. Lima: Less is more for alignment.

Tong Zhou, Yubo Chen, Pengfei Cao, Kang Liu, Jun Zhao, and Shengping Liu. 2023b. Oasis: Data curation and assessment system for pretraining of large language models.

A Author Component

A.1 Human Excerpts Extraction Algorithm

Algorithm 1 outlines the process for extracting human excerpts from movie scripts. The method involves maintaining a minimum window size, checking how many authors are within it, and determining if the dynamic window could be expanded further. Initially, if the starting window contains more than two actors, it is shifted by one unit; otherwise, it is enlarged until it meets a third actor, at which point the excerpt without the third actor is saved. After saving, the window returns to its minimum size and put in such a way that there cannot be overlapping turns with already saved dialogues. A final check (lines 19-20) is performed to save the last portion of an excerpt if it meets the specified criteria. The implementation of SAVE_SUB_DIALOGUE can be customized as required. For our use case we utilized a minimum window size of 3.

Algorithm 1 Dynamic Sliding Window

1: **function** DYNAMICSLIDINGWIN-DOW($scene, window\_size$)
2:     $begin \leftarrow 1$
3:     $end \leftarrow window\_size$
4:     **while** $end \leq$ (length of $scene$) **do**
5:         **if** UNIQUE($scene.author[begin : end]$)$> 2$ **then**
6:             $end \leftarrow end - 1$
7:             **if** $end - begin < window\_size$ **then**
8:                 $begin \leftarrow begin + 1$
9:                 $end \leftarrow begin + window\_size$
10:             **else**
11:                 SAVE_SUB_DIALOGUE($scene[begin : end]$)
12:                 $begin \leftarrow end + 1$
13:                 $end \leftarrow begin + window\_size$
14:             **end if**
15:         **end if**
16:         $end \leftarrow end + 1$
17:     **end while**
18:     $end \leftarrow end - 1$
19:     **if** UNIQUE($scene.author[begin : end]$)$= 2$ **and** $end - begin \geq window\_size$ **then**
20:         SAVE_SUB_DIALOGUE($scene[begin : end]$)
21:     **end if**
22: **end function**

A.2 Prompt for Dial$_{H+LLM}$ generation

Dial$_{H+LLM}$ has been generated by prompting Chat-GPT with the input specified in Table 11. While we present the prompt in English for ease of understanding, we utilised its Italian version in the actual interactions. The dialogues were generated using top-p decoding with a value of 0.9 and a temperature of 0.8 via OpenAI APIs.

PROMPT
Rewrite the following dialogue in its entirety in a way that is meaningful, natural, realistic, coherent, comprehensible and self-conclusive.
{{DIALOGUE}}

Table 11: Prompt to rewrite excerpts extracted from movies with ChatGPT.

A.3 Prompt for Dial$_{LLM}$ generation

To identify the most effective instruction, we conduct preliminary experiments involving testing various prompts. Annotators evaluated the quality of the generated data using a sample, based on factors like variability, originality, and coherence of the generations. In one experiment, we instructed the model to generate dialogues solely using contextual information. In another, we instructed it to first establish the personalities of two speakers based on the context, then generate the entire dialogue considering both the context and the speakers' personalities. We observed that explicitly requesting distinct personalities resulted in more varied dialogues, whereas without such instruction, the generated dialogues were similar for style and structure. Therefore, we provided the model with a specific context and instructed it to develop the personalities of two speakers engaging in a dialogue about that context. An example of prompt is provided in Table 12. Although we display the prompts in English for clarity, the actual interactions took place in Italian. Also in this case, the dialogues were generated using OpenAI APIs, employing *top-p* decoding with parameters set to 0.9 and a temperature of 0.8.

B Reviewer Component: Detailed Guidelines

In this section, we outline the methodology guidelines employed to review our dataset's dialogues. The main focus is to post-edit the Italian dialogue so to ensure that they appear natural, self-contained, and clear without being vague while minimising the manual intervention. Each dialogue must consist of at least three turns, with exactly two speakers taking alternating turns.

Optimal Dialogues and Turns

- Preferred dialogues and turns should be natural, realistic, contextually appropriate, and easily understandable.

Table 12: Example of prompt used to generate dialogues with ChatGPT, translated from Italian.

- **Italian dialects**: Italian dialects can be included briefly, but the main message should still be clear to readers unfamiliar with those dialects. Since Italian is mostly based on Tuscanian, this dialect is generally more tolerated.

- **Non-fluent expressions**: expressions such as *Ha*, *Ah* and *Oh* can be kept, although commonly associated with spoken language.

- **Expression related to communication mediums:** this kind of expressions, such as speaking into a radio, can be kept.

- **Common typos**: typos that are commonly associated with keyboard limitations are acceptable. However, if the same typos consistently appear in all turns, it is recommended to correct them occasionally (e.g. using *E'* instead of *È*).

Sub-optimal Dialogues and Turns

- **Dialogues with less than three turns**: if a dialogue consists of fewer than three turns, either the dialogue can be deleted or new turns can be added to ensure it has at least three turns;

- **Excluding non-Italian languages**: if a movie script excerpt contains passages in languages other than Italian, or if ChatGPT occasionally generates non-Italian dialogues, those non-Italian dialogues or turns should be removed. This rule also applies in-universe (e.g. fantasy) languages.

- **Extensive use of Italian dialects**: if a dialogue uses Italian dialects extensively, it becomes harder to read and understand. Therefore, such dialogue should be removed.

- **One-way communication**: This refers to a situation where one party in a conversation is merely providing verbal or non-verbal cues of attention or approval, without contributing any meaningful context or information. These dialogues should be post-edited or removed.

- **Distorted Italian**: Some characters might speak Italian in a distorted manner, influenced by their native language or origins. This distortion could involve using incorrect grammar, speaking only in infinitives, or mixing in excessive jargon from their original language (e.g."ogre speaking"). Dialogues or turns with this characteristic should be post-edited for clarity or, if necessary, removed.

- **Non-Italian onomatopoeic expressions**: onomatopoeic expressions not commonly found in Italian, such as *Euh*, can be removed.

- **Excessively vague dialogues**: sometimes, dialogues can be overly vague, lacking clear meaning or purpose. This typically occurs in short exchanges. Such dialogues should be removed.

- **Sung Parts:** Some dialogues contain sung sections, usually in English and recognizable by the use of the "/" character. These sections should be removed.

- **Individual speech**: sometimes, even if in a dialogue, characters might speak without replying to what the other is saying. This occurs, for instance, when characters are in conflict or when there is a speaker device like a radio or television involved. These instances should either be removed or adjusted during post-editing.

- **Non-verbal interruptions**: these interruptions occur in spoken dialogues when one

speaker uses non-verbal cues to signal agreement or encourage the other speaker to continue, resulting in the interruption of the current speaker's turn. These interruptions can be eliminated and the turns of the interrupted speaker can be merged.

- **Grammatical errors**: when a dialogue or a turn contains grammatical errors (e.g. dì instead of di' for meaning *dire*, *qual'è* instead of *qual è*, *qualcun'altro* instead of *qualcun altro*), they must be fixed.

- **References to external context**: If external content is frequently mentioned, especially through expressions like *qui* or *là*, consider removing such references.

- **Excessive politeness in dialogues**: Sometimes, ChatGPT generates dialogues that are excessively polite, with frequent agreement and continuous expressions of gratitude, resulting in unnatural conversations. Unnecessarily polite phrases can be removed. If an entire dialogue is excessively polite, it may be deleted.

How to perform turn deletion

Dialogues must consist of two alternating speakers. Therefore, any deletions in the middle of the conversation must be done in pairs. This means either removing one speaker's turn and merging it with the previous or next turn of that same speaker, or removing two turns together. The only exception to the pair deletion rule is when the turn slated for deletion is at the beginning or end of the dialogue. In such cases, a deletion may involve only one turn.

When a speaker in a dialogue delivers two consecutive turns, those turns should be merged into a single turn.

How to perform turn addition

Adding turns to a dialogue can significantly improve its clarity and coherence, particularly when there is a lack of context. We recommend only to add turns at the beginning or at end of a dialogues, as adding them midway requires a comprehensive understanding of the context prior and after the section in which is planned an addition, beyond the need to design an even number of turns so to keep the alternation of speaker, requirement which are considered excessively demanding effort-wise.

How to modify dialogues by merging or splitting

If there are two consecutive dialogues that appear as a continuation of each other, these can be merged.

In case there is a dialogue with multiple occurring contexts (e.g. when characters are discussing one topic and then something significant happens that shifts the focus). If both parts still meet the basic criteria of a valid dialogue (at least 3 exchanges that are clear and self-conclusive), they can be separated.

C Experimental Details

C.1 Fine-Tuning Configuration

Each Pythia model dimension was fine-tuned employing QLoRA (Dettmers et al., 2023), with low-rank approximation set to 64, low-rank adaptation set to 16, and dropout rate set to 0.1 Evaluation steps were set at 720, batch size at 4 and gradient accumulation step at 2. We employed 50 epochs with early stopping after 5 epochs. The original learning rates from Biderman et al. (2023) were utilised. The final epochs and learning rates are detailed in Table 13. For the fine-tuning process, an Nvidia Ampere A40 GPU with 48GB memory was utilised.

Model	Epochs	Learning Rate
1.4B$_{orig}$	11	2.0×10^{-4}
1.4B$_{p\text{-}e}$	10	2.0×10^{-4}
6.9B$_{orig}$	10	1.2×10^{-4}
6.9B$_{p\text{-}e}$	9	1.2×10^{-4}
12B$_{orig}$	19	1.2×10^{-4}
12B$_{p\text{-}e}$	13	1.2×10^{-4}

Table 13: Models' fine-tuning epochs and learning rates

C.2 Automatic Results Details

C.2.1 CPPL Details

Table 14 displays CPPL values with 20% and 30% of the last turns in dialogues progressively excluded. It demonstrates that models trained on post-edited data consistently outperform those trained on original data, even when the last turns of the dialogues are removed. Generally, removing the last turns improves model performance in terms of CPPL, indicating a negative impact of these turns on model performance.

Model	All turns	-20%	-30%
1.4B$_{orig}$	14.016	13.543	14.188
1.4B$_{p\text{-}e}$	11.554	11.369	11.638
6.9B$_{orig}$	13.098	10.511	11.011
6.9B$_{p\text{-}e}$	**10.465**	8.301	8.474
12B$_{orig}$	11.733	10.024	10.422
12B$_{p\text{-}e}$	10.985	**8.064**	**8.215**

Table 14: Models' CPPL values over gold turns, with 20% and 30% of dialogues' last turns progressively removed.

D   Human Evaluation

Example of dialogue negatively evaluated (Understandability of 1 and Machine Prob. of 4.667):

IT	EN
A: Ciao Giulia, come stai? Spero che tu possa aiutarmi in questa situazione difficile.	Hi Giulia, how are you? I hope you can help me in this difficult situation.
B: Ciao Luca, certo, dimmi pure qual è la tua situazione attuale.	Hi Luca, sure, please tell me what is your current situation.
A: Grazie per essere qui con me, mi hai raccontato tutto quello che sto cercando di fare.	Thank you for being here with me, you have told me everything I am trying to do.
B: Non c'è bisogno di spiegarmi, sei un grande amico.	No need to explain, you are a great friend.
A: Lo so, grazie ancora Giulia.	I know, thank you again Giulia.
B: È stato un piacere parlare con te. Buona giornata!	It's been a pleasure talking with you. Have a great day!
A: Grazie, a presto!	Thank you, see you soon!
B: Ciao!	Bye!

Table 15: Example of a negatively evaluated dialogue.

Example of dialogue positively evaluated (Understandability of 4.667 and Machine Prob. of 1.333):

IT	EN
A: Ciao Carla, come va?	Hi Carla, how are you doing?
B: Bene grazie, e tu?	Fine thank you, how about you?
A: Tutto bene, grazie. Volevo chiederti una cosa.	I'm all right, thank you. I wanted to ask you something.
B: Certo, dimmi pure.	Sure, tell me.
A: Vorrei sapere se ti piacerebbe venire a cena con me domenica prossima.	I would like to know if you would like to have dinner with me next Sunday.
B: Sì, sarebbe fantastico!	Yes, that would be great!
A: Perfetto, allora ci vediamo domenica prossima.	Perfect, then I'll see you next Sunday
B: Fantastico! A domenica!	Great! See you on Sunday!
A: A domenica!	See you on Sunday!

Table 16: Example of a positively evaluated dialogue.