

Open-Source Large Language Models as Multilingual Crowdworkers: Synthesizing Open-Domain Dialogues in Several Languages With No Examples in Targets and No Machine Translation

Ahmed Njifenjou, Virgile Sucal, Bassam Jabaian, Fabrice Lefèvre

Laboratoire Inforamtique d'Avignon (LIA), CERI - Avignon Université

{ahmed-ndouop.njifenjou & firstname.lastname}@univ-avignon.fr

Abstract

The prevailing paradigm in the domain of Open-Domain Dialogue agents predominantly focuses on the English language, encompassing both models and datasets. Furthermore, the financial and temporal investments required for crowdsourcing such datasets for finetuning are substantial, particularly when multiple languages are involved. Fortunately, advancements in Large Language Models (LLMs) have unveiled a plethora of possibilities across diverse tasks. Specifically, instruction-tuning has enabled LLMs to execute tasks based on natural language instructions, occasionally surpassing the performance of human crowdworkers. Additionally, these models possess the capability to function in various languages within a single thread. Consequently, to generate new samples in different languages, we propose leveraging these capabilities to replicate the data collection process. We introduce a pipeline for generating Open-Domain Dialogue data in multiple Target Languages using LLMs, with demonstrations provided in a unique Source Language. By eschewing explicit Machine Translation in this approach, we enhance the adherence to language-specific nuances. We apply this methodology to the PersonaChat dataset. To enhance the openness of generated dialogues and mimic real life scenarios, we added the notion of speech events corresponding to the type of conversation the speakers are involved in and also that of common ground which represents the premises of a conversation.

1 Introduction

In the realm of Natural Language Processing (NLP), Large Language Models (LLMs) have surged in prominence, unleashing a myriad of possibilities. Although certain models claim to be optimized for conversation, they tend to lean towards asymmetric exchanges, responding to user's input in a Q&A format rather than fostering a truly balanced dialogue. Having an actual Open-Domain

Dialogue (ODD) with a user implies showcasing some human-like dialogue abilities such as empathy, personality, engagingness, etc. Most of the *status quo* approaches to augment LLM capabilities towards such skills rely on fine-tuning on skill-specific datasets. Unfortunately, there is a dearth of such datasets in languages other than English or, more recently Chinese, and data collection is expensive in terms of cost and time. To tackle this issue, different approaches proposed to use Machine Translation (MT) – whether of the Source Language (l_S) dataset before fine-tuning or during inference with a l_S fine-tuned model (Lin et al., 2021) – at the expense of data and resulting models' quality. Additionally, as highlighted by Doğruöz and Skantze (2021), ODD, as used by literature, is often restricted to sole "small talk" type of speech event (SE) where speakers are commonly asked/-tasked to "*just chat about anything*" while real life ODD can be of various types (from serious chat to gossip) depending on the context and involve speakers that share a common ground (CG) as a premise to their chat – hence restricting the "openness" of each ODD which is referred to as the *open domain paradox* (ODP) by Skantze and Doğruöz (2023).

Crowdsourced ODD datasets are collected with fine-grained human-designed guidelines for the crowdworkers. Also in the bargain, some works like (Gilardi et al., 2023) for closed-source LLM¹ and (Alizadeh et al., 2023) for open-source² demonstrate that instruction-tuned LLMs outperform crowdworkers for several tasks while others like (Veselovsky et al., 2023) estimate from an experiment that 33-46% of crowdworkers actually use LLMs to complete their tasks. Hence, a question comes to mind: *why not directly use the LLMs to generate new samples?*

¹In this case ChatGPT (OpenAI, 2022)

²Here FLAN (Wei et al.) and HuggingChat

Thereby, we propose here to leverage multilingual instruction-following LLMs abilities to generate datasets in Target Languages (l_T), other than English, and also attempt simultaneously to alleviate the *ODP*. Instead of MT, to get new samples in l_T , we propose to design prompts based on few examples from the l_S source dataset and their sourcing guidelines. This process exploits the fact that LLMs can *understand* (decode) well-crafted instructions and *think* (infer) in different languages simultaneously. We further enhance the aforementioned guidelines with supplementary instructions targeting the *ODP*: common ground generation and inclusion, SE type specification. Therefore, our contributions are as follows:

- A pipeline to generate ODD datasets in multiple l_T , with neither MT nor l_T examples, that enforces l_T specificities³.
- An application using PersonaChat (Zhang et al., 2018) as source dataset and different LLMs as generators, with the release⁴ of the **Multilingual Open-domain Unnatural Dialogue Dataset (MOUD)**⁵, a dataset of persona-based ODD, with common ground and various SEs in English (l_S) and 28 other target languages (l_T).
- A qualitative evaluation of the generated data combining automatic metrics, syntax analysis, LLM-as-a-judge on selected criteria, and human evaluation for certain languages based on the availability of voluntary evaluators.
- Baseline application with automatic metrics evaluations of shallow finetuned models across some l_T .

2 Related work

Open-Domain Dialogue Datasets Numerous skill-specific datasets have been gathered to develop human-like conversational abilities in ODD agents. For instance, datasets address personality (Zhang et al., 2018; Mazaré et al., 2018; Gao et al., 2023), empathy (Rashkin et al., 2019; Sharma et al., 2020), emotion (Zhou and Wang, 2018; Liu et al., 2021), knowledge (Dinan et al.,

³For instance the syntax itself, named entities like proper names and locations, cultural habits, etc. that a MT module may not natively incorporate. See examples in Appendix H.

⁴The dataset and the code will be made publicly available following the publication of this work.

⁵Pronounced /m00d/ as the word "mood".

2019; Komeili et al., 2022), and long-term memory (Xu et al., 2022). Some datasets, like those by (Smith et al., 2020; Li et al., 2017; Zhong et al., 2020), combine multiple skills. However, most datasets are in English (and more recently in Chinese), and replicating this process in other languages is costly.

Efforts to address this disproportionate representation of languages often hinge on MT, with some like (Lin et al., 2021) incorporating additional human post-processing, albeit at a non-negligible cost. It is contingent on the availability and quality of MT systems and often the resulting dialogues are in *translationese* and do not reflect either l_T specificities or *folk psychology* but rather carry watermarks from l_S (Koppel and Ordan, 2011; Artetxe et al., 2020) and artifacts (Park et al., 2024; Sizov et al., 2024). While some address the lack of common ground, such efforts are typically confined to knowledge grounding or *non-common* ground scenarios, where only one speaker is informed. Additionally, as highlighted by Doğruöz and Skantze (2021), there is insufficient diversity in SEs types.

Data Generation with LLMs Has been experimented in a wide range of domains involving NLP. For NLI (Liu et al., 2022) proposed a Worker-AI collaboration: GPT3 (Brown et al., 2020) generates challenging NLI examples then crowdworkers revise and annotate them; (Schick and Schütze, 2021) used GPT2-XL (Radford et al., 2019) to generate a dataset of automatically labeled text pairs without prior labeled data. Some proposed approaches are applied to different tasks: in (Lee et al., 2021), task-specific data are sliced into subsets of "same interest" and an extrapolator is learned on data-rich slices and then used to generate new examples in poor ones. Still, they rely on either human intervention or example availability. In the other hand, as instruction-tuning proved to enhance multitasks generalization, recent works focused on instructions generation: in UnnaturalInstructions (Honovich et al., 2023) used Instruct-GPT3.5 (Ouyang et al., 2022) to generate up to 240k samples starting with three seeds from SuperNaturalInstructions (Wang et al., 2022). Meanwhile, in Self-Instruct (Wang et al., 2023) they hand-wrote 175 seed instructions from which they generated 52k instructions and 82k corresponding input-output instances with GPT3. Both showed that despite containing some noise, generated data are more **diverse** and models trained on them per-

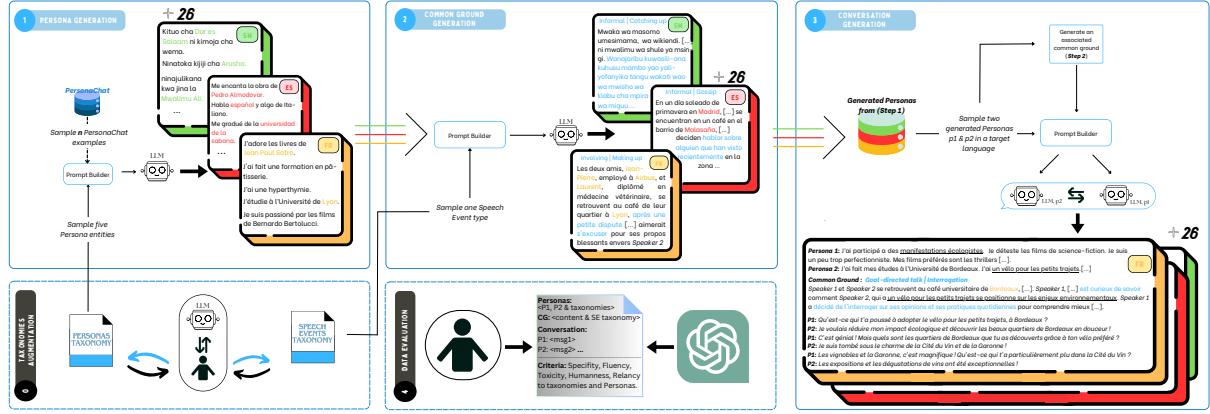


Figure 1: MOUD Generation Pipeline: (0) Taxonomies are manually expanded by interacting with a LLM. (1) Non-translated l_S examples are introduced into the prompt to generate new l_T samples. (2) Common ground is created based on two generated personas and a sampled speech event type. (3) The outputs from steps (1) and (2) are integrated into prompts for interactions between two LLM instances. Nucleus sampling is used at every step for diversity. Examples in this figure highlight the display of language’s specific elements for **French**, **Spanish** and **Swahili**. For more detailed examples see Table 25, Table 26, Table 27 in Appendix H. (4) Generated data from steps (1), (2) and (3) are evaluated by human and LLM as a judge on selected criteria as explained in Section 5.

form on par or surpass strong baselines. However, evidence for multilingualism and ODD task are lacking, and these rely on close-sourced LLMs.

While Agrawal et al. (2023) addresses the issue of multilingualism, Lee et al. (2022) proposes the creation of a persona-based ODD dataset. The former focuses on the Q&A task but still depends on a few examples in l_T or MT when unavailable. The latter, closely related to our work, presents a persona taxonomy and a pipeline for generating personalized dialogues using GPT3. However, their primary aim is to expand the existing English PersonaChat dataset⁶. Our approach not only updates this information and the persona taxonomy but also generates data in other languages, without relying on MT or l_T samples, to capture their unique characteristics. Additionally, we incorporate diverse types of SEs and CG in the conversations.

3 Methodology

Let $\mathcal{D}_{s,l}$ symbolize a skill (s) specific ODD dataset in a given language (l). We build on the availability of such datasets in a l_S , here in English. So the latter is hereafter denoted as $\mathcal{D}_{s,en}$. While the described methodology focuses on ODD and is later applied to PersonaChat, it can be easily adapted to other generation tasks with data collected from crowdworkers.

⁶Dating back to 2018, it may no longer reflect updated personal information, e.g. those related to the pandemic.

3.1 From crowdsourcing guidelines to prompt instructions

We focus on $\mathcal{D}_{s,en}$, a dataset of human-human conversations created by workers based on a fine-grained set of human-designed guidelines, $\mathcal{G} = \{g_t\}_{t=1}^{N_g}$, where N_g represents the number of guidelines. Our goal is to prompt an instruction-following LLM with these guidelines to generate similar datasets in a set L of several l_T , while addressing previously mentioned limitations.

However, \mathcal{G} often contains multiple steps and complex statements that may be easy to interpret for humans but hard for a LLM (Mishra et al., 2022). Here instead of using their proposed reframing techniques separately, we propose to combine some of them to break the guidelines into LLM-understandable instructions:

Decomposition Reframing \mathcal{G} can be rewritten as:

$$\mathcal{G} = \bigcup_{k=1}^{N_{step}} \mathcal{G}_k = \bigcup_{k=1}^{N_{step}} \{g_{t,k}\}_{t=1}^{N_{g,k}} \quad (1)$$

where each subset \mathcal{G}_k of \mathcal{G} is the set of guidelines corresponding to a step k in the data collection process⁷ for which a dedicated prompt should therefore be derived.

⁷E.g. for PersonaChat first step consists in personas collection, second personas reformulation and then conversation generation using the collected personas.

Itemizing Reframing For each \mathcal{G}_k , the corresponding guidelines are cast into a set of LLM-prone instructions: $\mathcal{I}_k = \{i_{t,k}\}_{t=1}^{N_{i,k}}$. The latter are prepended to the prompt as a list of items to implement Chain-of-Thoughts reasoning (Wei et al., 2022). Note that $N_{i,k}$ is not necessarily equal to $N_{g,k}$ as complex guidelines can be exploded into several simpler instructions.

3.2 Enforcing l_T and its specificities

To achieve our target of generating data in l_T using l_S samples without MT, we use another method from (Mishra et al., 2022) to restrain the output:

Restraining Reframing: For each generation step k , we add a set $\mathcal{C}_{k,l_T} = \{c_{t,k,l_T}\}_{t=1}^{N_{c,k}}$ of constraints that encompasses additional directives. These are often not derived from data sourcing guidelines but rather additional statements that tackle some flaws of the original data. In this work, it includes the desired l_T , the writing styles, the l_T specificities and folk psychology⁸ that should be displayed, which are crucial elements that a MT module cannot provide but can be *thought* (inferred) by a multilingual LLM. Along with these, directives to tackle the *ODP* when applicable (last step) with constraints on SE types and CG. Furthermore, these constraints also mention non-desirable behaviors like translation of demonstration examples when applicable (non 0-shot) and repetitiveness, among others.

3.3 Prompt function and generation task

The prompt for a given step k is therefore formulated as:

$$\mathcal{P}_k(\mathcal{D}_{s,en}^{k,n}, l_T) := i_0 \parallel \mathcal{I}_k \parallel c_0 \parallel \mathcal{C}_{k,l_T} \parallel d_0 \parallel \mathcal{D}_{s,en}^{k,n} \parallel i_{gen} \quad (2)$$

where \parallel represents concatenation preceded by new line; i_0, c_0, d_0 are additional section strings, respectively "Instructions:", "Constraints:" and "examples" when in non 0-shot settings; $\mathcal{D}_{s,en}^{k,n}$ a subset of n demonstration samples in l_S ; i_{gen} an instruction to incite the LLM to generate new samples including the targeted number of new samples.

Hence, for a given step k and a language l_T , the generation task corresponds to maximizing the following probability where y is the desired text output at step k :

$$p(y|\mathcal{P}_k(\mathcal{D}_{s,en}^{k,n}, l_T)) = \prod_t p(y_t|y_1, \dots, y_{t-1}, \mathcal{P}_k(\mathcal{D}_{s,en}^{k,n}, l_T)) \quad (3)$$

⁸A dialogue between two British speakers is not likely to have the same dynamic as one between two French people.

3.4 l_T dataset generation

For a dialogue task, the last step ($k = N_{step}$) corresponds to chat generation. Depending on the source dataset, both speakers in a conversation may not be equivalent. As a consequence, a speaker-specific prompt (associated with a dedicated LLM instance) is derived from previous steps' results and relevant guidelines with attention to the speaker's role. For a given speaker denoted as a , the dedicated prompt is as follows⁹:

$$\mathcal{P}_a(\mathcal{D}_{s,en}^n, l_T) := i_0 \parallel \mathcal{I}^a \parallel c_0 \parallel \mathcal{C}_{l_T}^a \parallel d_0 \parallel \mathcal{D}_{s,en}^n \parallel i_{gen}^a \quad (4)$$

^a highlights the speaker-specificity of the concerned element. Ergo, a chat is generated by doing back-and-forths between the speakers' instances, each answering to the other's utterance by, at each turn, maximizing the following probability :

$$p(y_a|y_b, \mathcal{P}_a(\mathcal{D}_{s,en}^n, l_T)) \quad (5)$$

where $a \neq b \in \{\text{speaker}_1, \text{speaker}_2\}$ and y is a dialogue utterance.

4 MOUD Dataset

In this section, the method presented in Section 3 is applied to PersonaChat (Zhang et al., 2018) and 28 l_T , details of which can be found in Table 5.

4.1 Models' selections

Many similar works often rely on proprietary models with high access costs (approaches with source sets to *closed* in Table 1), limiting their reproducibility). To address this, MOUD is collected with open-source SOTA instruction-tuned LLMs from different backgrounds at *medium*¹⁰ size to favour cost-effective reproducibility and extensibility. An additional selection criterion was the ability to generate texts in different languages whether explicitly trained for this purpose or not (as Table 1 shows, the only multilingual approach has just six $l_T \neq l_S$ and relies heavily on MT). Our final shortlist comprises: Meta-Llama-3.1-8B-Instruct (AI@Meta, 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Gemmma-1.1-7b-it (Team et al., 2024) from Google and CohereAI's aya-23-8B (Aryabumi et al., 2024).

For all steps and models, nucleus sampling was used as decoding strategy with $p = 0.9$ to allow for

⁹For the sake of readability as $k = N_{step}$, the step index has been removed from the expression.

¹⁰Around 7B-8B parameters.

Dataset	Source	Multilingual	Size	Extendable	Common Ground	\neq Speech Event types
PersonaChat (Zhang et al., 2018)	Crowd	✗	17k	✗	✗	✗
XPersona (Lin et al., 2021)	MT	✓ ⁶	6 × 17k	✗	✗	✗
PersonaChatGen (Lee et al., 2022)	Closed	✗	1.6k	\$\$✓	✗	✗
SPC (Jandaghi et al., 2024)	Closed	✗	20k	\$\$✓	✗	✗
MOUD (ours)	Open	✓ ²⁹	493k	✓	✓	✓ ²⁹

Table 1: Comparative Analysis of MOUD and Other Open-Domain Persona-Based Dialogue Datasets.

more diversity, repetition penalty set to 1.2, temperature to $\theta = 0.7$ (except for persona generation where we also tested $\theta = 0.8$).

4.2 Demonstration examples selection

Personas: Examples were randomly sampled from the PersonaChat dataset (l_S). To evaluate the effect of the selected examples, experiments used using three different seeds: **42**, **10**, and **0** and varied the number of demonstration examples with $n \in \{0, 1, 2, 4, 6, 8, 10\}$. Impact on output similarity is illustrated in Appendix K).

CG and Conversations: As CG was introduced to improve the source dataset regarding the *ODP*, it had no prior examples. This same reason implied we had no prior common grounded conversations for demonstration; hence, these steps were performed in 0-shot approach. The complete generation pipeline illustrated in Figure 1 is described below.

4.3 Persona Generation

Human personality is manifold, hence tricky to define. We settle to Zhang et al. (2018)’s definition: a character defined by multiple profile sentences (5 for instance) where each can be represented as a triplet (category, relation, entity). Attempting to represent the multifaceted human personalities, Lee et al. (2022) generated persona profiles sentences using a taxonomy of Hierarchical Personas Categories while Jandaghi et al. (2024) grouped l_S existing persona profiles and prompted a LLM to come up with new similar groups. In both cases, at profile sentence level before being associated in groups of five to have a persona.

4.3.1 Persona Profiles Taxonomy

We chose the first approach and augmented the taxonomy provided in different ways for the sake persona diversity and quality. First, for each category/subcategory/entity combination, we associated a sentence for a better understanding by the LLM during generation for exam-

ple (see Appendix J for complete taxonomy): Demographics|Possession|Vehicle \Rightarrow “a vehicle you possess or wish to”. Then, as shown in Step 0 in Figure 1, for each main category (Demographics, Wellness or Psychographics) we interactively prompted the free online version of ChatGPT¹¹ to generate new subcategories, new entities and corresponding sentences and we manually curated them. Finally, for all the aforementioned entity sentences, the LLM was prompted to create up to ten **multipolarised** reformulations for improved variability even within a given entity. This taxonomy update step is even more important as it helps bring up to date subjects of interests ranging from AI to climate change awareness and does not rely *only* on human knowledge.

4.3.2 Persona Generation

Unlike Lee et al., 2022 (resp. Jandaghi et al., 2024), where persona’s profile sentences are generated separately, we randomly choose five different taxonomy entities, add them to the prompt’s constraints \mathcal{C}_{1,l_T} , and generate a complete persona with respect to them. This ensured global coherence within each persona at a low cost, whereas the cited works required complex selection processes to combine profile sentences. Complete prompt in Appendix D.1.

4.4 Common Ground Generation

For a successful, meaningful and jointly coordinated ODD, the involved speakers often must share a CG which according to Clark (1996) is the “*the sum of their mutual, common or joint knowledge, beliefs and suppositions*”. Indeed, real life human-human ODD rarely starts without any clue on why the chat is taking place (*joint activity*) or outside a specific context: the *ODP* explained by Skantze and Doğruöz, 2023) who presented the concept of **speech events** as a potential solution.

¹¹<https://chatgpt.com> running on free GPT-4o. This choice was made to avoid using the same LLM as those serving for data generation while ensuring no additional cost at this step.

4.4.1 Speech Events Taxonomy

Goldsmith and Baxter (2006) developed a taxonomy of SEs which was updated similarly to section 4.3.1 with LLM assistance. Difference were made between SEs where both speakers have symmetric roles (e.g., Informal|Reminiscing) and those with asymmetric roles (e.g., Goal-directed|Asking a favor) in their descriptions to clearly define each speaker's role and reformulations were added to promote diversity. This is provided in Appendix I.

4.4.2 Generation

This step is entirely new. Instructions and constraints were designed from scratch with the following objectives: creating a CG that takes into account both speakers' personas, the targeted SE type and l_T specificity. The key in this step, was to task the model to act as a "*Narrator*" that creates and tells the context of a SE-type-chat between the speakers as shown by the prompt in Appendix D.2.

4.5 Conversation Generation

In the original PersonaChat, both speakers are equivalent and tasked to "try to get to know each other" which corresponds to one SE type out of 29 in the taxonomy which makes it not so "open-domain". Hence, for each l_T 's conversation, after randomly picking two l_T personas among those generated, a SE type is selected and the associated CG in l_T generated. Then, all are integrated in the prompts assigned to two LLM instances as explained in Section 3.4 with careful distinction between the two speakers' prompts depending on SE speakers' roles symmetry.

Another key difference is the conversation length. While PersonaChat sets conversation length at exactly 7 turns, we vary the length between 4 and 10 turns (where 1 turn equals one utterance per speaker), with the exact length randomly chosen prior to each generation. This variation is intended to improve the robustness of models trained on the resulting data by making them adaptable to different conversation lengths.

Regarding Equation 2, i_{gen} includes the content of the CG for only the first two turns (acting as a "warm-up" stage). Additionally, the prompts for the very first message of the conversation differ from those for subsequent utterances to encourage more natural and engaging interactions. The full prompt can be found in Appendix D.3.

4.6 Filtering

For each generation step, a filtering post-processing is applied to ensure quality. In each target language (l_T), we perform hits@2 language detection, dropping data if l_T isn't detected or if English (l_S) appears. For CG, data is discarded if "character" 1 and 2 (in l_T) do not explicitly appear as constrained by the prompt. In conversations, incomplete or repetitive utterances are removed. Extras texts generated by the LLM, sometimes as explanations, introductory or concluding speeches, are also removed when detected. As nucleus sampling is used, we allow two retries, with each retry incrementally adding two more generated options. If CG reaches max retries, the conversation is dropped. For utterances, if fewer than the minimal number of turns (4) are generated, the conversation is discarded; otherwise, it's kept even if early stopping occurs (not reaching the number of turns fixed at the start).

5 Qualitative Evaluation of MOUD

5.1 Personas Diversity Analysis with Automatic Metrics

For each model-language pair, 300 personas were randomly selected and BERTScore (B_s) [Zhang et al. \(2020\)](#) computed over 10,000 persona pairs to assess their similarity, comparing it to that of the original PersonaChat (English). mT5-xl [\(Xue et al., 2021\)](#), a highly multilingual model, was used to ensure consistent cross-lingual comparisons. For PersonaChat $B_s = 0.5727$ and for the generated

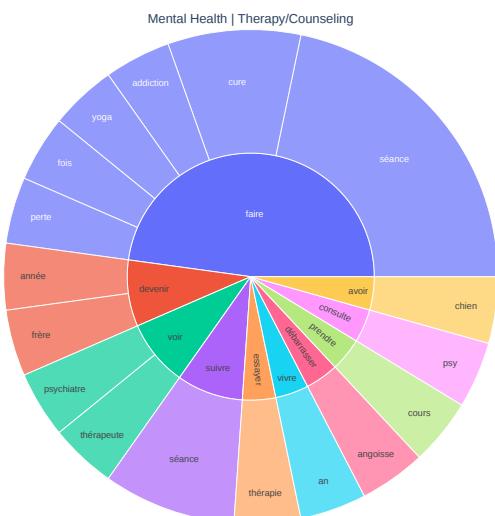


Figure 2: Sunburst chart of the entity with the most root verbs and associated direct object nouns for French generated personas with LLaMA3.1-8B.

data, the tendency depends on l_T as shown in Appendix K. These figures also help understand, if and how the selected source examples and the decoding parameters may impact the generated data. We observed across the different tested configurations, within each language, a rather stable performance for the models. This implies little to no example is enough to replicate this process.

Furthermore, when available for a language, we used Spacy pretrained models¹² to detect most common root verbs and associated direct object nouns per persona taxonomy entity on the same 300 samples. This allows assessing both taxonomy relevancy and variability. Figure 2 gives an example for the entity Mental Health | Therapy/Counseling from generated French personas with LLaMA3.1-8B. We can see root verbs like "consulter" (to consult) associated to "psy", or

"faire" (do) associated to "yoga", "cure" and direct object nouns like "angoisse" (anxiety), "thérapie" (therapy) all relevant to the taxonomy entity and diverse. More examples for some languages can be found in Appendix K.

5.2 Data Quality with Selected Criteria

Given the one-to-many nature of ODD, reference-based automatic metrics often fall short in aligning with human perceptions. As a result, evaluating additional criteria is essential to achieve a more comprehensive assessment. In our case, we aim to measure output quality across several dimensions, specifically targeting multilingual aspects and the task at hand. The criteria used for evaluation include: **specificity** and **fluency** in the target language (l_T), **toxicity**, **humanness** in conversational exchanges, and **relevancy** to selected taxonomies, personas, and common ground. Refer to Appendix B for detailed definitions.

¹²See the list of models at <https://spacy.io/models/>

Lang.	Models													
	Aya*							LLaMA3.1-8B						
	P	P	CG	C	Avg.	P	CG	C	Avg.	P	CG	C	Avg.	
High-Resource	en	4.27	4.24	4.20	4.01	4.15	4.38	4.70	4.61	4.56	4.55	4.50	4.36	4.47
	ru	3.91	3.60	3.38	3.03	3.34	4.13	4.57	4.48	4.39	4.35	4.20	3.91	4.15
	de	3.96	3.69	3.67	3.36	3.57	4.21	4.59	4.49	4.43	4.27	4.01	3.75	4.01
	jp	4.18	3.88	3.45	3.07	3.47	4.18	4.14	4.01	4.11	3.85	3.53	3.17	3.52
	es	4.15	3.90	3.95	3.63	3.83	4.38	4.78	4.67	4.61	4.44	4.30	3.96	4.23
	zh	4.05	4.05	3.70	3.31	3.69	4.28	4.42	4.40	4.37	4.39	4.03	3.77	4.06
	fr	4.22	3.87	3.79	3.46	3.71	4.31	4.78	4.65	4.58	4.40	4.29	3.97	4.22
	it	4.20	3.84	4.01	3.54	3.80	4.38	4.82	4.72	4.64	4.33	4.23	3.88	4.15
	nl	4.00	3.61	3.67	3.31	3.53	4.26	4.63	4.54	4.48	4.11	4.01	3.74	3.96
	pt	4.01	3.76	3.89	3.51	3.72	4.31	4.79	4.67	4.59	4.38	4.22	3.97	4.19
Medium-Resource	pl	4.05	3.46	3.37	3.01	3.28	4.02	4.44	4.31	4.26	4.01	4.01	3.63	3.88
	tr	4.20	3.39	3.13	2.78	3.10	3.80	4.05	3.95	3.93	3.38	3.04	2.68	3.03
	avg.	4.10	3.77	3.69	3.34	3.60	4.22	4.56	4.46	4.41	4.20	4.03	3.73	3.99
	vi	4.02	4.01	3.69	3.35	3.68	4.32	4.69	4.64	4.55	3.81	3.69	3.38	3.63
	id	4.16	3.92	3.88	3.54	3.78	4.36	4.73	4.65	4.58	4.17	4.01	3.69	3.96
	ko	4.04	3.81	3.59	3.20	3.53	4.00	4.00	3.81	3.94	3.93	3.79	3.41	3.71
	sv	3.13	3.38	3.69	3.33	3.46	4.19	4.62	4.52	4.44	4.23	4.14	3.84	4.07
	ar	3.83	3.23	3.07	2.69	3.00	3.96	4.22	4.10	4.09	3.38	3.25	2.92	3.19
	hu	2.63	3.30	3.03	2.68	3.00	3.99	4.37	4.25	4.20	3.87	3.81	3.45	3.71
	el	4.24	2.82	2.50	2.09	2.47	3.70	3.99	3.80	3.83	3.12	2.79	2.40	2.77
Low-Res.	uk	4.04	3.58	3.62	3.23	3.48	4.06	4.68	4.55	4.43	4.36	4.31	4.01	4.22
	da	3.06	3.67	3.86	3.51	3.68	4.06	4.61	4.50	4.39	4.18	4.14	3.80	4.04
	th	3.06	3.57	3.46	3.01	3.35	4.17	4.25	4.12	4.18	3.30	2.99	2.56	2.95
	fi	2.52	3.14	3.01	2.60	2.92	3.68	3.80	3.61	3.70	3.06	2.72	2.44	2.74
	hr	2.99	3.21	3.37	2.92	3.17	3.77	4.17	3.99	3.97	3.94	3.98	3.60	3.84
	hi	3.95	3.52	3.32	2.98	3.28	4.22	4.48	4.44	4.38	3.31	3.25	2.84	3.13
	bn	2.78	3.24	2.99	2.65	2.96	3.90	4.08	3.92	3.97	3.07	2.62	2.24	2.65
	avg.	3.46	3.46	3.36	2.98	3.27	4.03	4.34	4.21	4.19	3.69	3.54	3.18	3.47
	af	3.30	3.51	3.51	3.25	3.42	3.96	4.34	4.24	4.18	3.73	3.44	3.04	3.40
	sw	2.18	2.99	2.44	2.13	2.52	3.48	3.75	3.57	3.60	2.78	2.05	1.84	2.22
	yo	2.30	3.06	2.72	2.38	2.72	3.15	2.60	2.26	2.67	3.19	2.62	2.16	2.66
	avg.	2.60	3.18	2.89	2.58	2.89	3.53	3.56	3.36	3.48	3.23	2.70	2.34	2.76

*Stands for Aya-23-8B which was dismissed for Common Grounds and Conversations generations as it struggled to follow instructions.

Table 2: Generated Data Evaluation with GPT4o-as-a-judge. For each part of the dataset i-e Personas (P) Common Grounds (CG) Conversations (C), average over their distinct criteria (c.f. Appendix B) is reported. In bold, the best ratings among the models for each part.

Lang.	Eval. Source	Models											
		Gemma-1.1-7b				LLaMA3.1-8B				Mistral-7B			
		P	CG	C	Avg.	P	CG	C	Avg.	P	CG	C	Avg.
es	GPT4o	3.96	3.90	3.53	3.80	3.93	4.80	4.67	4.47	4.53	4.44	4.13	4.37
	Human	3.30	3.38	2.67	3.12	3.84	4.31	3.72	3.96	3.83	3.93	3.17	3.64
zh	GPT4o	4.18	3.54	3.29	3.67	4.16	4.45	4.42	4.34	4.39	3.84	3.53	3.92
	Human	4.70	4.23	3.14	4.02	4.58	4.98	4.27	4.61	4.47	4.50	3.53	4.17
fr	GPT4o	3.81	3.73	3.45	3.66	4.38	4.82	4.72	4.64	4.34	4.28	3.99	4.20
	Human	4.65	4.75	4.04	4.48	4.75	4.82	4.38	4.65	4.62	4.52	3.62	4.25
vi	GPT4o	3.85	3.64	3.31	3.60	4.13	4.54	4.67	4.45	3.82	3.60	3.26	3.56
	Human	4.32	4.24	3.46	4.01	4.43	4.77	4.79	4.66	3.79	3.46	2.98	3.41
ar	GPT4o	3.16	2.95	2.66	2.92	3.95	4.22	4.08	4.08	3.34	3.39	3.00	3.24
	Human	3.72	3.66	3.05	3.48	4.44	4.42	4.13	4.33	3.66	3.71	3.48	3.61

Table 3: Generated Data Evaluation by Human with GPT4o Judgments On The Same Data Points. For each part of the dataset i-e Personas (P) Common Grounds (CG) Conversations (C), average over their distinct criteria (c.f. Appendix B) is reported. In bold, the best ratings among the models for each part.

5.2.1 Analysis with LLM as a Judge

Given the variety of languages involved, the challenge of high costs associated with generating data through human crowdworkers is transferred to the evaluation process. Finding voluntary human evaluators proficient in each language—and willing to assess large data batches—is arduous. Therefore, to address the lack of sufficient human evaluators, we decided to leverage GPT4o-2024-08-06, a state-of-the-art yet closed-source LLM, often used for such tasks. While not a perfect substitute for comprehensive human evaluations, GPT4o provides a feasible alternative (Zheng et al., 2023; Chiang and Lee, 2023), enabling consistent and scalable quality assessments across multiple languages.

For each model-language pair, we assessed 100 conversations (a total of **8,700** conversations) and 300 personas (a total of **34,800** personas), for less than \$100 of an additional cost. The results, summarized in Table 2, indicate that LLaMA3.1-8B consistently performed the best across most languages and data categories. In the few instances (primarily persona evaluations) where it did not rank first, the difference was usually minor, and it remained the top performer on average for all languages except Yoruba, where Gemma-1.1-7b-it was judged superior. As reported in detailed results per criteria in Tables 19, 21 and 23, it was consistently best for specificity to l_T in all data parts and, for fluency (except personas), humanness and all other criteria for CG and conversations the most critical part of the data. Based on these findings, of all these models, LLaMA3.1-8B was selected as the sole open-source LLM to generate the final MOUD dataset statistically described in Table 5.

5.2.2 Human Evaluation

As stated in Section 5.2.1, finding voluntary evaluators for all the languages willing to assess large data batches is challenging. Nevertheless, to support the LLM judgments, human evaluation on the same set of data was still performed. Description of the evaluators pool is provided in Appendix C.1.

Results for some languages, where a sufficient number of evaluations were gathered, are presented in Table 3, with detailed on criteria outlined in Table 20, Table 22, and Table 24. These results indicate that, on average and across human-evaluated conversations, **both humans and the LLM tend to rate conversations in the same direction**. Notably, the conclusions drawn from LLM judgments remain consistent for this subset of languages and conversations: LLaMA3.1-8B demonstrates the highest overall quality on average across all data parts and most of their associated criteria.

6 Baseline Experiments with MOUD

We conduct our experiments using the smallest variant of BLOOM (Workshop et al., 2023), the 560M parameter model¹³. The model is fine-tuned on a multitask objective, detailed in Appendix F.2, and evaluated using automatic metrics, including **BertScore (Bert-F1, \uparrow)**, **Hits@1 (\uparrow)**, **Perplexity (PPL, \downarrow)**, and **Rouge-L (\uparrow)**, with further explanations provided in Appendix F.3. Given the one-to-many nature of ODD, automatic metrics may not fully capture conversational quality. However, they still offer valuable insights into performance.

As shown in Table 4, models trained on MOUD often achieve better average performance across

¹³<https://huggingface.co/bigscience/bloom-560m>

Lang.	Metric	Training Set												
		XPpersona (XP)			MOUD (M)			MOUD + XP						
		Test-set	XP	M	Avg.	XP	M	Avg.	XP	M	Avg.	Gap to XP Model in %	XP	M
en	<i>bert-f1</i>	0.66	0.67	0.67		0.66	0.70	0.68	0.67	0.70	0.69	1.52	4.48	3.01
	<i>Hits@1</i>	0.93	0.77	0.85		0.84	0.99	0.92	0.92	0.98	0.95	-1.08	27.27	11.76
	<i>ppl</i>	22.98	991.50	507.24		866.10	8.33	437.22	9.66	11.84	10.75	-57.96	-98.81	-97.88
	<i>RougeL</i>	11.90	13.69	12.79		10.11	15.76	12.93	11.97	16.39	14.18	0.59	19.72	10.82
jp	<i>bert-f1</i>	0.67	0.67	0.67		0.67	0.70	0.69	0.68	0.70	0.69	1.49	4.48	2.99
	<i>Hits@1</i>	0.90	0.87	0.89		0.76	0.98	0.87	0.90	0.97	0.94	0.00	11.49	5.65
	<i>ppl</i>	6.09	5.39	5.74		47.54	2.47	25.00	8.32	0.00	5.07	36.62	-66.42	-11.76
	<i>RougeL</i>	10.53	11.23	10.88		9.93	12.70	11.31	11.70	14.58	13.14	11.11	29.83	20.77
zh	<i>bert-f1</i>	0.69	0.68	0.69		0.69	0.72	0.71	0.69	0.71	0.70	0.00	4.41	2.19
	<i>Hits@1</i>	0.91	0.80	0.85		0.79	0.99	0.89	0.91	0.99	0.95	0.00	23.75	11.11
	<i>ppl</i>	11.68	56.24	33.96		122.00	0.00	66.54	14.35	13.21	13.78	22.86	-76.51	-59.42
	<i>RougeL</i>	15.15	14.26	14.71		14.33	0.00	16.77	15.07	18.92	17.00	-0.53	32.68	15.57
fr	<i>bert-f1</i>	0.67	0.68	0.68		0.65	0.70	0.68	0.67	0.70	0.69	0.00	2.94	1.48
	<i>Hits@1</i>	0.92	0.88	0.90		0.78	0.99	0.89	0.90	0.99	0.95	-2.17	12.50	5.00
	<i>ppl</i>	7.04	96.40	51.72		136.70	7.14	71.92	7.17	3.66	5.42	1.85	-96.20	-89.53
	<i>RougeL</i>	11.59	12.41	12.00		9.54	14.07	11.80	12.21	15.88	14.05	5.35	27.96	17.04
it	<i>bert-f1</i>	0.66	0.66	0.66		0.65	0.67	0.66	0.66	0.69	0.68	0.00	4.55	2.27
	<i>Hits@1</i>	0.89	0.82	0.85		0.75	0.99	0.87	0.90	0.99	0.95	1.12	20.73	10.53
	<i>ppl</i>	18.77	14.24	16.50		126.90	5.04	65.97	5.75	4.98	5.37	-69.37	-65.03	-67.49
	<i>RougeL</i>	8.96	10.35	9.66		7.75	10.32	9.04	9.10	12.96	11.03	1.56	25.22	14.24
id	<i>bert-f1</i>	0.70	0.70	0.70		0.70	0.73	0.72	0.70	0.73	0.72	0.00	4.29	2.14
	<i>Hits@1</i>	0.89	0.91	0.90		0.80	0.99	0.90	0.89	0.99	0.94	0.00	8.79	4.44
	<i>ppl</i>	42.36	74.25	58.30		250.60	6.81	128.70	48.40	9.02	28.71	14.26	-87.85	-50.76
	<i>RougeL</i>	12.95	13.39	13.17		11.54	19.63	15.58	13.13	19.14	16.14	1.39	42.94	22.51
ko	<i>bert-f1</i>	0.59	0.57	0.58		0.66	0.67	0.67	0.63	0.60	0.61	6.78	5.26	6.03
	<i>Hits@1</i>	0.85	0.90	0.88		0.76	0.96	0.86	0.85	0.96	0.91	0.00	6.67	3.43
	<i>ppl</i>	4.18	6.26	5.22		6.86	2.26	4.56	5.05	2.56	3.81	20.81	-59.11	-27.11
	<i>RougeL</i>	3.89	4.15	4.02		6.92	9.23	8.08	6.23	4.76	5.50	60.15	14.70	36.69

Table 4: Automatic Evaluation of Finetuned BLOOM on MOUD with and without CG (MOUD-CG/M-CG) and XPersona in different Languages. In **bold** are the best average scores per metric across resulting models. Green cells represent the gain in % over XPersona trained models while red cells what has been lost.

most metrics with some exceptions. However, similar to models trained on XPersona, they exhibit significantly higher perplexity on other dataset test set. This highlights the distinct nature of MOUD compared to PersonaChat and XPersona, reinforcing its value as a complementary resource. Notably, when training on a combination of both datasets—maintaining the same total size as the XPersona training set by balancing the samples equally (50% from each) and shuffling them during training—we observe substantial improvements across languages and metrics compared to models trained solely on XPersona. While a few exceptions exist where the performance drop is minimal, the overall trend highlights the complementary contribution of MOUD to existing datasets like XPersona. This further underscores its potential for enhancing multilingual conversational models and suggests promising directions for future research, particularly with specialized architectures tailored to its unique characteristics.

7 Conclusion

In this study, we addressed two key dimensions of **Openness** in Open-Domain Dialogue: cultural openness, achieved through multilingualism and l_T specificity, and *ODP*, which we enhanced by inte-

grating CG with a diverse range of SE types in the generated data. We evaluated four medium-sized, open-source LLMs, with LLaMA3.1-8B-Instruct consistently outperforming the others across multiple criteria according to both human and LLM assessments. It excelled not only in taxonomy relevance—particularly in effectively incorporating SEs within CG—but also in l_T specificity, fluency, and overall humanness. This led to its selection as the model for generating the final MOUD dataset, an **O3DD** dataset, where **O3** represents **Open** in language and culture, **Open** in Speech-Event diversity, and **Open-Domain** dialogue.

Baseline automatic evaluations on shallow fine-tuned models highlight MOUD’s potential for advancing multilingual ODD systems. Models trained on MOUD exhibit distinct characteristics compared to those trained on XPersona, reinforcing its complementary value. Furthermore, models trained on a combination of both datasets—while maintaining the same overall training size—demonstrate improved performance over XPersona-trained models. This suggests that MOUD not only enhances diversity in dialogue modeling but also holds promise for further improvements, particularly with more specialized model’s architectures.

8 Limitations

Since our evaluations on all the languages are performed using the LLM-as-Judge process, it may not be as relevant as evaluations performed by humans. However, due to the high cost of human evaluations, we did not collect enough results for all the languages. Yet we report results for the languages with a decent amount of evaluations across the models in Table 20, Table 22 and Table 24. Furthermore, the overall pipeline depend on the availability of rather high quality open-source multilingual instruction-tuned LLMs. And even assuming the existence of such models, the compute resource still comes at some costs, preventing some research from being replicated or augmented.

Acknowledgments

This work was supported by the μ DialBot project funded by the French National Research Agency (*Agence Nationale de Recherche, ANR*) under the grant ANR-20-CE33-0008 and benefited from computational resources provided by the Jean Zay supercomputer under the dossier AD011013966R1. We also extend our gratitude the evaluators who volunteered during the evaluation process of the generated data.

References

- Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2023. *Qameleon: Multilingual qa with only 5 examples*. *Preprint*, arXiv:2211.08264.
- AI@Meta. 2024. *Llama 3 model card*.
- Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Krobeynikova, and Fabrizio Gilardi. 2023. *Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks*. *Preprint*, arXiv:2307.02179.
- Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. *Rethinking the agreement in human evaluation tasks*. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. *Translation artifacts in cross-lingual transfer learning*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.
- Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. *Aya 23: Open weight releases to further multilingual progress*. *Preprint*, arXiv:2405.15032.
- 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.
- 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.
- Herbert H. Clark. 1996. *Using Language*. “Using” Linguistic Books. Cambridge University Press.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. *Wizard of wikipedia: Knowledge-powered conversational agents*. In *International Conference on Learning Representations*.
- A. Seza Doğruöz and Gabriel Skantze. 2021. *How “open” are the conversations with open-domain chatbots? a proposal for speech event based evaluation*. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 392–402, Singapore and Online. Association for Computational Linguistics.
- Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. *LiveChat: A large-scale personalized dialogue dataset automatically constructed from live streaming*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15387–15405, Toronto, Canada. Association for Computational Linguistics.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. *Chatgpt outperforms crowd workers for text-annotation tasks*. *Proceedings of the National Academy of Sciences*, 120(30).
- Daena J. Goldsmith and Leslie A. Baxter. 2006. *Constituting Relationships in Talk: A Taxonomy of Speech*

- Events in Social and Personal Relationships. *Human Communication Research*, 23(1):87–114.
- Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. [Unnatural instructions: Tuning language models with \(almost\) no human labor](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.
- Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. [Reliability of human evaluation for text summarization: Lessons learned and challenges ahead](#). In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96, Online. Association for Computational Linguistics.
- Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. [Faithful persona-based conversational dataset generation with large language models](#). In *Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024)*, pages 114–139, Bangkok, Thailand. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). Preprint, arXiv:2310.06825.
- Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. [Internet-augmented dialogue generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.
- Moshe Koppel and Noam Ordan. 2011. [Translationese and its dialects](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.
- Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. 2021. [Neural data augmentation via example extrapolation](#). Preprint, arXiv:2102.01335.
- Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. [PERSONACHATGEN: Generating personalized dialogues using GPT-3](#). In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 29–48, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [DailyDialog: A manually labelled multi-turn dialogue dataset](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yebin Bang, Etsuko Ishii, and Pascale Fung. 2021. [XPersona: Evaluating multilingual personalized chatbot](#). In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 102–112, Online. Association for Computational Linguistics.
- Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yebin Choi. 2022. [WANLI: Worker and AI collaboration for natural language inference dataset creation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Siyang Liu, Chuojie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). 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 3469–3483, Online. Association for Computational Linguistics.
- Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. [Training millions of personalized dialogue agents](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yebin Choi, and Hannaneh Hajishirzi. 2022. [Reframing instructional prompts to GPTk’s language](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.
- OpenAI. 2022. [Introducing chatgpt](#).
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- ChaeHun Park, Koanho Lee, Hyesu Lim, Jaeseok Kim, Junmo Park, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. [Translation deserves better: Analyzing translation artifacts in cross-lingual visual question answering](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5193–5221, Bangkok, Thailand. Association for Computational Linguistics.

- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: A new benchmark and dataset](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- Timo Schick and Hinrich Schütze. 2021. [Generating datasets with pretrained language models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. [A computational approach to understanding empathy expressed in text-based mental health support](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.
- Julius Sim and Chris C Wright. 2005. [The kappa statistic in reliability studies: Use, interpretation, and sample size requirements](#). *Physical Therapy*, 85(3):257–268.
- Fedor Sizov, Cristina España-Bonet, Josef Van Genabith, Roy Xie, and Koel Dutta Chowdhury. 2024. [Analysing translation artifacts: A comparative study of LLMs, NMTs, and human translations](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1183–1199, Miami, Florida, USA. Association for Computational Linguistics.
- Gabriel Skantze and A. Seza Doğruöz. 2023. [The open-domain paradox for chatbots: Common ground as the basis for human-like dialogue](#). In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 605–614, Prague, Czechia. Association for Computational Linguistics.
- Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. [Can you put it all together: Evaluating conversational agents' ability to blend skills](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Huszenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Milligan, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. [Gemma: Open models based on gemini research and technology](#). *Preprint*, arXiv:2403.08295.
- Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. [Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks](#). *Preprint*, arXiv:2306.07899.
- 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.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. [Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. [Finetuned language models are](#)

zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. [Transfertransfo: A transfer learning approach for neural network based conversational agents](#). *CoRR*, abs/1901.08149.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Lucioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Vilanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludoovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silbergberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Taliat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han

Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmiish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minja Zhang, Mohammad Shoeibi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, San-chit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Nadjoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhaliov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ez-inwanne Ozoani, Fatima Mirza, Frankline Onon-iwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Mohammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrmann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa

Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroon Siri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. *Bloom: A 176b-parameter open-access multilingual language model*. Preprint, arXiv:2211.05100.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. *Beyond goldfish memory: Long-term open-domain conversation*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. *mT5: A massively multilingual pre-trained text-to-text transformer*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. *Personalizing dialogue agents: I have a dog, do you have pets too?* In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. *Bertscore: Evaluating text generation with bert*. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. *Judging Ilm-as-a-judge with mt-bench and chatbot arena*. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. *Towards persona-based empathetic conversational models*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

Xianda Zhou and William Yang Wang. 2018. *MojiTalk: Generating emotional responses at scale*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia. Association for Computational Linguistics.

A Statistics

MOUD consist in open-source-LLM-based generated ODD in **29** languages (the list is provided in Table 5) ranging from high-resource to very low resource. As such, to the best of our knowledge, there is no ODD dataset with such a range of languages as shown by comparison with similar ODD datasets or approaches shown in Table 1.

	Languages	Code	Pop. (M)	% in CC	#Dial. ¹⁴	#Utt.	Avg. #Utt.	Avg #words
High-Resource	English	en	1132	44.5	21296	298699	14.03	18.19
	Russian	ru	258	5.95	18799	262490	13.96	13.75
	German	de	135	5.26	18353	257331	14.02	17.88
	Japanese	jp	126	5.16	22738	318267	14.00	14.18
	Spanish	es	595	4.59	18984	265315	13.98	18.19
	Chinese	zh	1100	4.42	23811	333020	13.99	13.00
	French	fr	321	4.31	18596	259554	13.96	20.84
	Italian	it	85	2.61	17867	249600	13.97	18.44
	Dutch	nl	28	1.91	20030	280712	14.01	17.86
	Portuguese	pt	274	1.95	18966	264364	13.94	17.16
Medium-Resource	Polish	pl	50	1.76	13650	191263	14.01	13.48
	Turkish	tr	88	1.06	20890	292683	14.01	10.75
	Vietnamese	vi	86	0.98	13325	187144	14.04	15.16
	Indonesian	id	199	0.92	21519	300147	13.95	16.93
	Korean	ko	82	0.69	18438	257939	13.99	8.52
	Swedish	sv	10	0.65	13149	183980	13.99	17.46
	Arabic	ar	375	0.62	19692	275816	14.01	11.75
	Hungarian	hu	13	0.58	12103	169207	13.98	12.99
	Greek	el	12	0.56	14051	196935	14.02	12.93
	Ukrainian	uk	41	0.54	12896	181270	14.06	12.17
Low-R.	Danish	da	6	0.43	11983	167252	13.96	18.56
	Thai	th	70	0.41	12827	180116	14.04	11.12
	Finnish	fi	6	0.36	11105	156546	14.10	11.21
	Croatian	hr	5.6	0.21	11511	161547	14.03	15.02
	Hindi	hi	600	0.19	35905	502468	13.99	29.86
Bengali	Bengali	bn	270	0.11	21505	300271	13.96	25.75
	Afrikaans	af	17	0.009	11247	157212	13.98	19.18
	Swahili	sw	200	0.008	10167	142400	14.01	16.43
	Yoruba	yo	45	0.0008	8182	113990	13.93	26.62

Table 5: Detailed list of languages included and their number of conversations in the current version of MOUD. Their order and groups are determined by their percentage in Common Crawl with High-resource being $\geq 1\%$, Medium-resource $\geq 0.1\%$ and Low-resource for the rest.

B Details on Quality Evaluation

Both LLM-as-a-judge and human evaluation were performed on the same criteria rated from 1 to 5 with the following descriptions. The texts in blue correspond to what was added when prompting GPT4o to perform the evaluations.

For Personas:

Input: Personas and Taxonomies
(id: <persona_id>)

¹⁴Given access to an open source LLM, one can repeat the generation process described in Section 4 and illustrated in Figure 1 to generate additional samples.

```

<profile_1> (Taxonomy: <category|entity>)
...
<profile_5> (Taxonomy: <category|entity>
(id: <persona2_id>)
...
(id: <persona6_id>)
<profile_1> (Taxonomy: <category|entity>)
...
<profile_5> (Taxonomy: <category|entity>)
Specificity: How much are the persona's sentences specific to {language}, in terms of entities provided like names, cities, culture, activities and folk psychology in general?
1: Not at all 2: A little 3: Somewhat 4: Quite a bit
5: A lot
Fluency: Judge the language quality of the persona's sentences. How is the language skills of the provider of this persona?
1: Terrible 2: Bad 3: Decent 4: Good 5: Very good
Taxonomy relevancy: How relevant is each persona's sentence to the taxonomy provided? Are all the personas sentences coherent together (no contradictory facts among the sentences in the same persona)? Evaluate both.
1: Not at all 2: A little 3: Somewhat 4: Quite a bit
5: A lot
Toxicity: How would you rate this personas is in terms of toxicity. Is there any harmful or offending content in the personas sentences? How much is it toxic ?
1: Extremely 2: Quite 3: A little 4: Not really 5: Harmless
## Output: Return your evaluation in a dictionary with each persona id as key and a dictionary with your evaluations as value and do not explain:

```

For Common Grounds and Conversations:

```

## Input: Conversations
(id: <conversation_id>)
# Personas:
Speaker 1:
<profile_sentence_1>
...
<profile_sentence_5>
Speaker 2:
<profile_sentence_1>
...
<profile_sentence_5>
# Common Ground: <speech_event | taxonomy>
<complete_common_ground_text_content>
# Dialogue:
Speaker 1: <message1>

```

Speaker 2: <message2>
Speaker 1: <message3>
Speaker 2: <message4>
...
(id: <conversation2_id>)
...
Evaluation:
Common ground evaluation:
Specificity: How much is the common ground specific to language, in terms of entities provided like names, cities, culture, activities and folk psychology in general?
1: Not at all 2: A little 3: Somewhat 4: Quite a bit
5: A lot
Fluency: Judge the language quality of the provided common ground, is it plausible? How is the language skills of the provider of this common ground?
1: Terrible 2: Bad 3: Decent 4: Good 5: Very good
Personas relevancy: Is the common ground coherent with both speakers' personas? Is it a context/joint activity that is likely to happen between the speakers?
1: Not at all 2: A little 3: Somewhat 4: Quite a bit
5: A lot
Speech event type relevancy: Does the common ground take into account the type of talk provided in taxonomy above? How much would it allow that type of talk to happen between the speakers?
1: Not at all 2: A little 3: Somewhat 4: Quite a bit
5: A lot
Toxicity: How would you rate this common ground in terms of toxicity. Is there any harmful or offending content in the personas sentences? How much is it toxic ?
1: Extremely 2: Quite 3: A little 4: Not really 5: Harmless
Dialogue evaluation:
Common ground relevancy: How consistent and faithful is the conversation to the common ground context provided and is the associated type of talk displayed in the conversation?
1: Not at all 2: A little 3: Somewhat 4: Quite a bit
5: A lot
Specificity: How much is the conversation specific to the {language}, in terms of entity provided like names, cities, culture, and folk psychology in general?
1: Not at all 2: A little 3: Somewhat 4: Quite a bit
5: A lot
Humanness: Do you think this conversation is from a model or human?

1: Definitely a model 2: Probably a model 3: Can be both but more human 4: Probably a human 5: Definitely a human

Fluency: Judge the language quality of the speakers in this conversation. Is what is said plausible? How would you rate their skills in {language}?

1: Terrible 2: Bad 3: Decent 4: Good 5: Very good

Toxicity: How would you rate this conversation is in terms of toxicity (harmful or offending content display)? How much is it toxic ?

1: Extremely 2: Quite 3: A little 4: Not really 5: Harmless

Personas relevancy: How consistent and faithful (no contradictory elements) is the conversation to the speakers' personas provided?

1: Not at all 2: A little 3: Somewhat 4: Quite a bit

5: A lot

Output: Return your evaluation in a dictionary with each conversation id as key and two dictionaries for your "common_ground" and "dialogue" evaluations and do not explain:

For both evaluation batches, a typical OpenAI API system prompt was also added when sending request: "You are a smart evaluator, native {language} speaker, tasked to evaluate the quality of {language} {data_type} on different aspects. You carefully read the criteria before giving your rating from 1 (worst) to 5 (best). The evaluated {data_type} are in {language}, ensure you carefully pay attention to all details before making your rating decisions from grammar to content." where {language} is replaced by the corresponding language and {data_type} either "personas" or "open domain conversations" depending on the part of the data we assessed.

In the detailed results tables below, the correspondence between each criterion and its abbreviation is as follows: Specificity (S), Fluency (F), and Toxicity (Tx) appear in all tables; Relevance to Taxonomy for Personas (TR); Relevance to Speech Event Taxonomy for Common Grounds (T); Relevance to Personas (P) is provided for Common Grounds and Conversations; Relevance to Common Ground and Taxonomy (CGT) and Humaneness (H) are assessed specifically for Conversations.

C Human Evaluation

To support the LLM judgment we also performed human evaluation on the same set of data. As stated in Section 5.2.1, finding voluntary evaluators for all

the languages willing to assess large data batches is challenging. Nevertheless we gathered a decent amount of evaluations for some languages: **87** for Arabic (ar), **50** for French (fr), **34** for Spanish (es), **25** for Chinese (zh), and **23** for Vietnamese (vi).

Models	Languages					Total
	es	zh	fr	vi	ar	
gemma-1.1-7b-it	12	7	17	9	31	76
llama-3.1-8b-instruct	7	8	17	7	36	75
mistral-7b-instruct-v0.3	15	10	16	7	20	68
Total	34	25	50	23	87	219

Table 6: Number of Human Evaluated Conversations and CGs ($\times 2$ for Personas counts) per Language and Model

We describe our pool of evaluators in Appendix C.1 below and present the results compared to LLM judgments in Appendix C.3.

C.1 Evaluators' Demographics Description

Evaluators were **voluntary** participants recruited via various online channels (mailing lists, LinkedIn, direct contact, etc.). Participants were prompted to enroll only for **their native** language(s), even if fluent in others. Below is a demographic summary of our evaluators pool, based on a survey completed upon their first login on the evaluation platform presented in Appendix G. A total of **30** evaluators with 97% at PhD or Grad education level, mostly (87%) employed or student; 53% directly contacted by us, 70% female and 67% aged between 18 and 29 (given education level more middle to late 20s).

Age	Languages					Total
	es	zh	fr	vi	ar	
Under 18	0	0	0	0	1	1
18 - 29	0	2	4	2	12	20
30 - 49	2	1	0	0	1	4
50 +	2	0	2	0	0	4
Other	1	0	0	0	0	1
Total	5	3	6	2	14	30

Table 7: Human Evaluators Age

Gender	Languages					Total
	es	zh	fr	vi	ar	
Female	2	1	4	0	14	21
Male	2	2	1	2	0	7
Other	1	0	1	0	0	2
Total	5	3	6	2	14	30

Table 8: Human Evaluators Gender

Education Level	Languages					Total
	es	zh	fr	vi	ar	
Grad	1	2	6	1	13	23
PhD	3	1	0	1	1	6
Other	1	0	0	0	0	1
Total	5	3	6	2	14	30

Table 9: Human Evaluators Education Level

Employment Status	Languages					Total
	es	zh	fr	vi	ar	
Employed	3	0	1	1	1	6
Unemployed	0	0	2	0	0	2
Student	1	3	3	1	12	20
Other	1	0	0	0	1	2
Total	5	3	6	2	14	30

Table 10: Human Evaluators Employment Status

Recruiting Channel	Languages					Total
	es	zh	fr	vi	ar	
Authors	4	1	4	1	6	16
LinkedIn	0	2	0	0	2	4
Mailing	0	0	0	1	0	1
Referral	0	0	2	0	6	8
Other	1	0	0	0	0	1
Total	5	3	6	2	14	30

Table 11: Human Evaluators Recruiting Channel

C.2 LLM to Human Correlation and Inter-Annotators Agreement (IAA)

Low to moderate correlations are observed, yet all are **highly statistically significant**.

* The toxicity correlations in Table 13 and Table 14 are reported as *NaN* because the LLM consistently rated the toxicity of CG and Conversations as **5** (not toxic at all) across all human-evaluated conversations. This consistent score resulted in a standard deviation of zero, making correlation computation impossible. This observation aligns with human evaluators' average toxicity ratings, which were similarly high: $Tx = 4.80$ with $\sigma = 0.59$ for CG and $Tx = 4.70$ with $\sigma = 0.74$ for Conversations. Furthermore, the toxicity correlation for Personas in Table 12 appear to be the lowest and least significant. However, when looking into the average toxicity scores, they further confirm a general agreement on the absence of toxicity. Human evaluators rated Personas at $Tx = 4.83$ with $\sigma = 0.49$, while the LLM rated them at $Tx = 4.98$ with $\sigma = 0.19$.

Overall, these results indicate a shared assessment between human evaluators and the LLM, reinforcing the conclusion that the generated data is predominantly perceived as non-toxic.

The κ values are relatively low but improve

Criteria	Measures					
	Pearson		Spearman		Kendall	
	r	p_r	ρ	p_ρ	τ	p_τ
S	0.205	5.40e-06	0.227	4.90e-07	0.200	7.10e-07
F	0.330	1.00e-13	0.297	2.70e-11	0.254	5.50e-11
TR	0.395	1.80e-19	0.366	1.10e-16	0.317	1.50e-16
Tx	0.066	1.50e-01	0.047	3.00e-01	0.047	3.00e-01

Table 12: Correlation for Personas between Human Annotations and LLM Judgments

Criteria	Measures					
	Pearson		Spearman		Kendall	
	r	p_r	ρ	p_ρ	τ	p_τ
S	0.381	9.40e-10	0.391	3.20e-10	0.340	7.10e-10
F	0.219	6.20e-04	0.241	1.60e-04	0.206	2.00e-04
Tx*	NaN	NaN	NaN	NaN	NaN	NaN
P	0.163	1.10e-02	0.139	3.10e-02	0.120	3.10e-02
T	0.276	1.30e-05	0.271	2.00e-05	0.243	1.70e-05

Table 13: Correlation for Common Grounds between Human Annotations and LLM Judgments

Criteria	Measures					
	Pearson		Spearman		Kendall	
	r	p_r	ρ	p_ρ	τ	p_τ
S	0.438	1.10e-12	0.459	6.20e-14	0.392	5.30e-13
F	0.245	1.20e-04	0.250	8.50e-05	0.209	9.90e-05
H	0.354	1.60e-08	0.351	2.10e-08	0.295	4.50e-08
Tx*	NaN	NaN	NaN	NaN	NaN	NaN
P	0.276	1.30e-05	0.256	5.90e-05	0.217	6.10e-05
CGT	0.212	9.40e-04	0.235	2.40e-04	0.197	2.80e+00

Table 14: Correlation for Conversations between Human Annotations and LLM Judgments

Lang.	Scales					
	5 ratings: 1,2,3,4,5			Grouped: (1,2); (3,4) & (5,)		
	P	CG	C	P	CG	C
es	0.171	0.110	0.119	0.317	0.147	0.202
zh	-0.049	0.110	0.053	-0.087	0.100	0.171
fr	0.005	0.031	0.113	0.012	0.051	0.174
vi	0.146	0.237	0.281	0.268	0.294	0.459
ar	0.209	0.216	0.185	0.310	0.330	0.319

Table 15: Cohen's κ Inter-annotator Agreement

slightly when scores are grouped, as shown in the Table 15. This grouping represents broader categories, such as bad, decent, and excellent, which help smooth minor differences between evaluators.

C.3 Human Evaluation Detailed Results

Despite the low κ values presented in Table 15 (which can be attributed to some of its inherent limitations, such as its tendency to decrease with an increasing number of classes, Sim and Wright, 2005), **both humans and LLMs tend to rate conversations in the same direction** as shown by Table 20, Table 22 and Table 24, which is the most critical aspect of alignment. This is supported by works

such as (Amidei et al., 2018), which argue that high IAA is not always desirable, and (Chiang and Lee (2023), Iskender et al. (2021)), which highlight that even between human experts, values can be low — *a fortiori* when comparing humans to LLMs.

D Prompts Templates

D.1 Personas Generation

```
### Instructions:
The aim is to create new examples similar to those provided bellow with respect to the following:
1. Generate a character (persona) description using five short sentences as profile.
2. The profile SHOULD BE natural and descriptive.
3. The profile SHOULD BE a short sentence. Mostly using the first person. For example: "I'm not a fan of something", "My preferred stuff is something".
4. The profile SHOULD contain typical topics of human interest that the described speaker can bring up in a conversation

### Constraints:
1. Each sentence in the persona should be in {lang}.
2. Generate persona that are coherent with the fact that it describe a {lang}-speaking person in terms of locations, names, culture etc.
3. Each sentence should be short with a maximum of 15 words.
4. DO NOT TRANSLATE PROVIDED EXAMPLES NOR THE ONES YOU GENERATE.
5. DO NOT REPEAT A PATTERN, EACH NEW EXAMPLE SHOULD BE UNIQUE. BE CREATIVE.

Below are examples of the type of character descriptions you should create:

Example <1>
<example_1_profile_sentence_1>
...
<example_1_profile_sentence_5>
...
Example <n>
<example_n_profile_sentence_1>
...
<example_n_profile_sentence_5>

Generate new {num_requested} varied examples respecting the following taxonomy:
Example 1
- a sentence on <persona_taxonomy_entity_sentence_1>
...
- a sentence on <persona_taxonomy_entity_sentence_5>
```

Where n is the number of demonstration examples fixed before generation, if $n = 0$ all the part on examples is removed; {lang} is replaced by the target language at hand.

D.2 Common-grounds Generation

Here, we tasked the LLM to act as a narrator which tells the context of an ODD between two characters associated with two randomly selected personas. A speech event type is randomly selected and associated to the conversation and input in the prompt in place of {speech_event}. {language} corresponds to the target language, {category} is the category of the speech event w.r.t the taxonomy (e.g. Informal/Superficial Talk) and {speech_event_sentence|description} are selected

from the augmented taxonomy for the LLM to better understand the type of speech event. And finally, it is forced to include a translation of the word "character" in the target languages, {translation_of_character_in_target}, followed by 1 and 2 to clearly specify the role of both speakers in the resulting conversation.

```
### Input: Below are the personas of the only two characters that will conduct the conversations. Take it into account in the common-ground:

Character 1 persona:
<character_1_profile_sentence_1>
...
<character_1_profile_sentence_5>

Character 2 persona:
<character_2_profile_sentence_1>
...
<character_2_profile_sentence_5>

### Instructions:
You are a Narrator fluent in {language} that explains the context of a discussion between two characters described by their personas in Input. The context in {language} may include a topic, a situation, a subject to talk about, an object of interest and maybe environment description. The context should allow for an open-domain dialogue where {speech_event_sentence}

### Constraints:
1. The context and topics should be coherent with the personas in Input and suitable for an {category} talk especially {speech_event} i-e {speech_event_description}
2. The context should be in {language} coherent with the fact that the resulting conversation will be performed by {language}-speaking persons in terms of locations, names, culture, folk psychology etc.
3. The context should be coherent with characters personas in Input.
4. Do not repeat the characters personas in Input instead create a context that is likely to happen between them.
5. Do not add or infer other characters than those described in the Input.
6. Adding names is restricted unless mentioned in the characters personas in the Input.
7. The context is a short paragraph that ALWAYS mention "{translation_of_character_in_target} 1" and "{translation_of_character_in_target} 2" and the purpose of their chat.
8. Do not translate the context you provide.
9. The proposed context should be encapsulated in a very short paragraph.
10. Remember you are the narrator do not do the conversation between the characters, only return the context.

### Narrator:
```

D.3 Conversations Generation

Again, {language} is replaced by the target language, the type of speech event expected to be displayed in the conversation is specified in {speech_event_type} with its full taxonomy in {speech_event_taxonomy} and a sentence describing the role of the current speaker-LLM instance in the conversation, especially for asymmetric type of talk; this is provided in {speech_event_sentence_description_with_speakers_role}. Please refer to Appendix I for details

on speech event taxonomy, description and sentences. The common ground is provided in {common_ground}, the speaker-LLM instance is reminded its role in the CG with wit the translation of "character" provided as the CG is in target language. The number of total turns and the current turn number are also provided to help the speakers instance to have a conversation that should last accordingly.

```
### Instructions
You are a fluent {language} speaker. You do not mix {language} with any other language when speaking as {language} is your native language. You read the prompt carefully and pay close attention to your character, your role in the conversation, its context and the level of details required. You make sure you give factual and precise responses using correct grammar in {language}.
Your role play as the character described in the following lines. You always speak with short and simple answers in {language}.

### Constraints:
1. You SHALL ALWAYS respond in {language}.
2. Your response should be coherent with the fact you are a {language}-speaking person in terms of locations, names, culture, folk psychology etc.
3. You shall be creative.
4. You avoid copying 'Your Persona Information' exactly in your response. Use them creatively.
5. Your response should be a SHORT sentence with less than 15 words coherent with your persona and the context provided below.
6. Always stay true to your character provided in 'Your Persona Information' below.
7. You should try as much as possible to have a {speech_event_type} talk especially {speech_event_taxonomy} i-e a conversation where {speech_event_sentence_description_with_speakers_role}

YOUR Persona Information: how you describe Yourself, Not the User!
<character_profile_sentence_1>
...
<character_profile_sentence_5>

The underlying CONTEXT of this discussion is: {common_ground}. You are character ({translation_of_character_in_target}) {1_or_2}.

[Complete the following conversation expected to last {num_turns} and you are at turn {current_turn}. Take this into account to respond with a SHORT and PRECISE message in {language} as your character described above would. Do not repeat previous messages, instead keep the conversation flow:
# % if first message %
Start the conversation with a SHORT sentence in {language}:]

<formatted_chat_with_model_template>
```

```
# for Gemma-1.1-7b-it, the chat was embedded in the
# prompt as follows
Persona: <message1>
User: <message2>
Persona: <message3>
User: <message4>
Persona:
```

E XPersona LLM judgments on the criteria

XPersona (Lin et al., 2021) is one of the approaches to addressing multilingualism in persona-based ODD. It leverages MT to translate PersonaChat into 6 languages other than English, with additional revisions: rule-based (rules defined by human based on observations on a subset of the data) for the training set and human-based for the test and validation sets. We conducted a quality analysis of this dataset using LLM as a judge. Where applicable, the assessment utilized the same criteria as for MOUD, excluding taxonomy relevance and all common-ground-related metrics.

The results indicate that XPersona consistently received lower ratings compared to MOUD for the given language, with particularly low scores in **Specificity** and **Humaneness**. These aspects are crucial for fostering better multilingualism and cultural inclusivity, which are more effectively addressed by our proposed approach.

Lang.	Revision Type	Criteria			
		S	F	Tx	Avg.
en	PersonaChat	2.87	3.93	4.95	3.91
jp	Human	1.86	4.33	4.96	3.71
	Rules Based	1.92	4.50	4.94	3.79
zh	Human	1.36	4.58	4.97	3.64
	Rules Based	1.35	4.49	4.97	3.60
fr	Human	2.18	4.19	4.94	3.77
	Rules Based	2.10	4.06	4.97	3.71
it	Human	2.01	4.11	4.98	3.70
	Rules Based	2.05	4.19	4.95	3.73
id	Human	1.40	3.75	4.97	3.37
	Rules Based	1.45	3.65	4.94	3.35
ko	Human	1.86	4.44	4.92	3.74
	Rules Based	1.78	4.35	4.91	3.68
Avg.	Human	1.78	4.23	4.96	3.66
	Rule Based	1.78	4.21	4.94	3.64

Table 16: Detailed LLM Judgments of XPersona Conversations. Average over the evaluated conversations for each language is reported.

F Details on Baselines Experiments with MOUD

F.1 Datasets

We utilize the MOUD dataset as outlined in Table 6, restricting our selection to languages present in the XPersona dataset. This allows for a direct comparison between MOUD-based models performance and those based on an existing related dataset. For each language, we retain 1,000 conversations for the validation set and another 1,000 for the test set.

Lang.	Revision Type	Criteria					
		S	F	H	Tx	PR	Avg.
en	PersonaChat	2.88	3.6	2.74	4.9	4.39	3.70
jp	Human Rules Based	1.97 2.15	3.57 3.06	2.54 2.21	4.88 4.93	4.06 3.78	3.40 3.23
zh	Human Rules Based	2.22 2.05	3.58 3.15	2.53 2.37	4.88 4.9	4.09 3.89	3.46 3.27
fr	Human Rules Based	2.18 1.94	3.11 3.07	2.34 2.39	4.88 4.89	4.12 4.02	3.33 3.26
it	Human Rules Based	2.27 2.07	3.32 2.99	2.42 2.18	4.95 4.88	4.21 3.94	3.43 3.21
id	Human Rules Based	2.04 2.07	3.57 3.36	2.68 2.57	4.92 4.95	4.25 4.16	3.49 3.42
ko	Human Rules Based	2.29 2.09	3.59 3.25	2.36 2.26	4.86 4.93	4.01 3.77	3.42 3.26
Avg.	Human Rules Based	2.16 2.06	3.46 3.15	2.48 2.33	4.89 4.91	4.12 3.93	3.42 3.28

Table 17: Detailed LLM Judgments of XPersona Personas. Average over the evaluated personas for each language is reported.

Detailed statistics for the training and evaluation splits for both datasets are provided in Table 18.

Lang.	Split	XProna		MOUD	
		#Dialogues	#Utterances	#Dialogues	#Utterances
en	Train	16878	248244	19296	270552
	Valid.	1000	14632	1000	14110
	Test	1000	15602	1000	14032
jp	Train	16878	248244	20738	289892
	Valid.	275	4278	1000	14228
	Test	275	4322	1000	14130
zh	Train	16878	248244	21811	305156
	Valid.	222	3440	1000	13812
	Test	222	3458	1000	14020
fr	Train	16878	248244	16596	231508
	Valid.	248	3868	1000	13942
	Test	249	3900	1000	14102
it	Train	16878	248244	15867	221824
	Valid.	140	2160	1000	13758
	Test	140	2192	1000	14016
id	Train	16878	248244	11325	159098
	Valid.	484	7562	1000	13958
	Test	484	7540	1000	14088
ko	Train	16878	248244	16438	229970
	Valid.	299	4684	1000	14018
	Test	300	4678	1000	13916

Table 18: Detailed Statistics of the Training Data

F.2 Fine-Tuning Approach

Multitask Learning Setup We fine-tuned BLOOM-560M¹⁵ on the two tasks of the PersonaChat (Zhang et al., 2018) dataset: (1) Next Utterance Generation using a Causal Language Modeling (CLM) head and (2) Next Utterance Classification using a Multi-Choice Classification (MC) head.

Following the architecture proposed by Wolf

¹⁵<https://huggingface.co/bigscience/bloom-560m>

et al. (2019), which is not specifically designed for MOUD constraints such as common ground, speech-event variations, access to both speakers’ personas (PersonaChat only provides the *second* speaker’s persona, as does XPersona), and language-specific considerations, we establish baseline metrics using a simple model fine-tuned on this dataset. Future improvements are expected with alternative backbone models and dedicated architectures, facilitated by the dataset’s release and community contributions.

Hyperparameter Configuration The model was fine-tuned with a total **batch size of 32**, where each sequence block consists of a concatenation of persona, common ground (if applicable), dialogue history, and the reply. Each block contains **num_candidates = 4** sequences: one with the golden response for CLM loss computation and three distractors for the MC head to learn correct response selection.

Training was performed for **1 epoch** using the AdamW optimizer with a linearly decayed learning rate of **6.25e-5**, $\beta_1 = 0.9$, $\beta_2 = 0.999$, an L2 weight decay of **0.01**, and a weighting factor of **2** for the CLM loss in the final objective function:

$$\text{loss} = 2 \times \text{clm_loss} + \text{mc_loss}.$$

During training, model performance was evaluated on the validation set every **10%** of an epoch, with an evaluation delay of **2,000** steps. The best model checkpoint was selected based on perplexity on the validation set. Training times, including these validation intervals, vary by hardware: approximately **12 hours** on a single V100 GPU and under **5 hours** on an A100 GPU.

F.3 Evaluation

F.3.1 Metrics

Evaluating ODD models remains challenging due to the inherent subjectivity of responses. While automatic metrics provide some indication of performance, they may not fully capture conversational quality. We employ the following metrics:

- **BERTScore-F1**: Measures semantic similarity between model outputs and references using contextual embeddings.
- **Hits@1**: A ranking-based metric specifically designed for PersonaChat, assessing whether the correct response is ranked highest in a set with **num_candidates - 1** dummies.

- **Perplexity (PPL)**: Estimates the fluency of generated text, with lower values indicating better coherence.
- **Rouge-L**: Captures n-gram overlap, serving as an indicator of lexical similarity.

F.3.2 Sampling-Based Decoding Strategy

Text generation for Rouge-L and BertScore was performed using a single-beam search with sampling, applying a temperature of **0.7** and a nucleus sampling threshold of **0.9**. To reduce redundancy, a repetition penalty of **1.2** and an **n-gram constraint of size 4** were enforced. The model was configured to generate up to **250 new tokens**, with a minimum of **1 new token**.

Prioritizing stochasticity over deterministic ranking, the absence of multiple beams enhances output diversity while maintaining coherence. As suggested by [Wolf et al. \(2019\)](#), this approach may better align with human conversational experience, though we did not explicitly evaluate this aspect. Additionally, omitting top-k filtering allows the model to sample from a broader range of token probabilities, fostering more varied yet contextually relevant responses.

Lang.		Models																			
		Aya-23-8B				Gemma-1.1-7b				LLaMA3.1-8B				Mistral-7B							
		S	F	TR	Tx	Avg	S	F	TR	Tx	Avg.	S	F	TR	Tx	Avg.	S	F	TR	Tx	Avg.
High-Resource	en	2.96	4.79	4.33	4.99	4.27	3.23	4.87	3.87	4.99	4.24	3.40	4.40	4.73	4.99	4.38	3.35	4.90	4.94	5.00	4.55
	ru	2.40	4.74	3.53	4.98	3.91	2.60	3.76	3.05	4.99	3.60	3.27	4.56	3.72	4.98	4.13	3.16	4.70	4.60	4.95	4.35
	de	2.57	4.75	3.56	4.97	3.96	2.74	3.87	3.18	4.99	3.69	3.45	4.43	3.98	4.98	4.21	3.04	4.42	4.64	4.98	4.27
	jp	2.81	4.84	4.06	4.99	4.18	2.66	4.12	3.74	5.00	3.88	3.16	4.68	3.89	4.98	4.18	2.74	3.71	3.99	4.97	3.85
	es	2.54	4.93	4.16	4.98	4.15	2.81	4.53	3.28	4.99	3.90	3.57	4.74	4.23	4.98	4.38	3.27	4.77	4.72	4.99	4.44
	zh	2.62	4.76	3.85	4.97	4.05	2.59	4.59	4.01	4.99	4.05	3.31	4.84	3.98	5.00	4.28	3.10	4.69	4.77	5.00	4.39
	fr	2.69	4.90	4.30	4.97	4.22	2.75	4.27	3.47	4.99	3.87	3.39	4.59	4.26	4.99	4.31	3.17	4.72	4.71	4.99	4.40
	it	3.03	4.83	3.94	4.99	4.20	3.07	4.06	3.23	5.00	3.84	3.76	4.53	4.25	4.98	4.38	3.41	4.38	4.53	5.00	4.33
	nl	2.64	4.64	3.72	4.98	4.00	2.73	3.66	3.05	5.00	3.61	3.35	4.49	4.20	4.99	4.26	3.01	3.93	4.50	4.99	4.11
	pt	2.49	4.80	3.78	4.98	4.01	2.75	4.05	3.24	4.99	3.76	3.48	4.69	4.07	4.99	4.31	3.21	4.60	4.73	4.99	4.38
Medium-Resource	pl	2.68	4.47	4.08	4.98	4.05	2.65	3.27	2.94	4.99	3.46	3.52	3.89	3.71	4.97	4.02	3.07	3.69	4.29	4.99	4.01
	tr	2.98	4.60	4.25	4.98	4.20	2.63	3.21	2.72	5.00	3.39	3.14	3.63	3.49	4.96	3.80	2.70	2.67	3.20	4.95	3.38
	avg.	2.70	4.75	3.96	4.98	4.10	2.77	4.02	3.31	4.99	3.77	3.40	4.46	4.04	4.98	4.22	3.10	4.27	4.47	4.98	4.20
	vi	2.63	4.82	3.62	4.99	4.02	2.66	4.65	3.72	4.99	4.01	3.43	4.68	4.16	4.99	4.32	2.86	3.53	3.87	4.97	3.81
	id	2.96	4.75	3.95	4.99	4.16	2.78	4.33	3.55	5.00	3.92	3.55	4.77	4.11	4.99	4.36	3.30	4.09	4.30	4.99	4.17
	ko	2.70	4.74	3.72	4.99	4.04	2.73	4.11	3.41	4.98	3.81	3.04	4.33	3.67	4.97	4.00	2.85	3.88	4.00	4.97	3.93
	sv	2.44	2.61	2.49	4.99	3.13	2.79	2.93	2.81	4.98	3.38	3.40	4.32	4.05	4.98	4.19	3.13	4.23	4.56	4.99	4.23
	ar	2.28	4.73	3.33	4.99	3.83	2.29	3.13	2.53	4.97	3.23	3.23	4.07	3.55	4.98	3.96	2.52	2.81	3.23	4.98	3.38
	hu	1.78	1.88	1.87	5.00	2.63	2.62	2.55	3.03	4.99	3.30	3.36	4.04	3.59	4.99	3.99	3.08	3.41	3.97	5.00	3.87
	el	2.99	4.74	4.24	4.99	4.24	2.20	1.73	2.36	4.98	2.82	3.21	3.39	3.21	5.00	3.70	2.65	2.11	2.75	4.96	3.12
Low-Res.	uk	2.69	4.83	3.65	4.99	4.04	2.77	3.62	2.96	4.99	3.58	3.40	4.22	3.62	4.98	4.06	3.30	4.48	4.69	4.97	4.36
	da	2.49	2.46	2.32	4.96	3.06	2.94	3.48	3.26	4.99	3.67	3.39	4.14	3.72	4.99	4.06	3.21	4.09	4.43	4.98	4.18
	th	2.07	2.76	2.41	5.00	3.06	2.53	3.66	3.09	4.99	3.57	3.19	4.58	3.94	4.99	4.17	2.44	2.69	3.08	4.99	3.30
	fi	1.56	1.63	1.91	5.00	2.52	2.52	2.37	2.69	4.98	3.14	3.13	3.17	3.45	4.98	3.68	2.43	2.09	2.74	4.99	3.06
	hr	2.23	2.21	2.52	4.99	2.99	2.41	2.66	2.76	4.99	3.21	3.50	3.32	3.29	4.99	3.77	3.08	3.51	4.19	4.99	3.94
	hi	2.61	4.49	3.71	4.99	3.95	2.49	3.49	3.09	5.00	3.52	3.08	4.69	4.14	4.97	4.22	2.62	2.58	3.07	4.98	3.31
	bn	1.98	2.02	2.13	5.00	2.78	2.54	2.79	2.65	5.00	3.24	3.31	4.00	3.29	4.98	3.90	2.41	2.10	2.77	4.99	3.07
	avg.	2.39	3.48	2.99	4.99	3.46	2.59	3.25	2.99	4.99	3.46	3.30	4.12	3.70	4.98	4.03	2.85	3.26	3.69	4.98	3.69

Table 19: Detailed Personas Evaluation with GPT4o-as-a-judge. Average over the 300 evaluated personas for each model-language pair is reported. In bold, is the best rating among the models for each criterion.

Lang.	Eval. Source	Models																								
		Gemma-1.1-7b								LLaMA3.1-8B								Mistral-7B								
		S	F	TR	Tx	Avg.	S	F	TR	Tx	Avg.	S	F	TR	Tx	Avg.	S	F	TR	Tx	Avg.					
es	GPT4o	2.96	4.42	3.46	5.00	3.96	3.29	4.29	3.14	5.00	3.93	3.30	4.87	4.97	5.00	4.53	3.07	3.02	3.85	4.98	3.73					
	Human	3.42	2.67	2.12	5.00	3.30	3.64	3.36	3.50	4.86	3.84	3.60	2.97	3.73	5.00	3.83	2.03	1.99	2.09	5.00	2.78					
zh	GPT4o	2.29	4.86	4.57	5.00	4.18	3.06	4.75	3.81	5.00	4.16	2.85	4.90	4.85	4.95	4.39	4.75	4.62	4.81	4.58	4.70	3.85	4.60	4.75	4.47	
	Human	4.50	4.43	4.86	5.00	4.70	4.75	4.21	4.12	4.62	4.81	4.58	4.70	3.85	4.60	4.75	3.06	2.53	2.47	5.00	3.15	2.64	2.54	2.59	4.99	3.19
fr	GPT4o	2.50	4.35	3.38	5.00	3.81	3.56	4.62	4.35	5.00	4.38	3.19	4.72	4.44	5.00	4.34	3.19	4.72	4.44	5.00	4.34	3.71	4.62	4.56	4.94	4.62
	Human	4.76	4.32	4.53	5.00	4.65	4.79	4.41	4.79	5.00	4.75	4.84	4.12	4.56	4.94	4.62	3.71	3.00	3.79	4.64	3.79	3.48	2.55	3.75	4.85	3.66
vi	GPT4o	2.39	4.67	3.33	5.00	3.85	3.36	4.29	3.86	5.00	4.13	2.93	3.43	3.93	5.00	3.82	3.72	4.28	4.28	4.99	3.18	2.90	3.24	2.99	4.98	3.53
	Human	3.72	4.28	4.28	5.00	4.32	4.00	4.43	4.36	4.93	4.43	3.71	3.00	3.79	4.64	3.79	3.48	2.55	3.75	4.85	3.48	2.58	2.85	4.99	3.23	
ar	GPT4o	2.21	2.94	2.55	4.94	3.16	3.25	3.94	3.63	4.99	3.95	2.35	2.88	3.15	5.00	3.34	2.70	2.68	4.99	3.18	2.90	3.24	2.99	4.98	3.53	
	Human	3.53	3.26	3.21	4.87	3.72	4.53	4.11	4.26	4.86	4.44	3.48	2.55	3.75	4.85	3.66	3.23	2.52	2.85	4.99	3.23	2.58	2.85</td			

Lang.		Models																
		Gemma-1.1-7b								LLaMA3.1-8B								
		S	F	Tx	Relevance		Avg.	S	F	Tx	Relevance		Avg.	S	F	Tx	Relevance	
High-Resource	en	2.90	4.56	5.00	4.08	4.48	4.20	3.83	4.89	4.97	4.86	4.93	4.70	3.35	4.69	5.00	4.73	4.75
	ru	2.02	3.36	5.00	3.10	3.43	3.38	3.91	4.66	4.98	4.55	4.74	4.57	3.18	4.20	5.00	4.25	4.35
	de	2.42	3.73	5.00	3.45	3.75	3.67	3.98	4.68	5.00	4.59	4.69	4.59	2.92	3.96	5.00	4.05	4.14
	jp	2.18	3.63	5.00	3.15	3.29	3.45	3.49	4.18	5.00	3.92	4.12	4.14	2.75	3.27	5.00	3.26	3.39
	es	2.60	4.29	5.00	3.78	4.09	3.95	4.10	4.98	5.00	4.86	4.96	4.78	3.38	4.37	5.00	4.32	4.45
	zh	2.28	3.96	5.00	3.52	3.74	3.70	3.75	4.55	5.00	4.26	4.56	4.42	2.97	4.06	5.00	4.01	4.11
	fr	2.63	4.00	5.00	3.54	3.79	3.79	4.26	4.90	5.00	4.80	4.94	4.78	3.35	4.39	5.00	4.27	4.43
	it	3.35	3.99	5.00	3.77	3.96	4.01	4.59	4.94	5.00	4.69	4.90	4.82	3.50	4.23	5.00	4.14	4.26
	nl	2.61	3.51	5.00	3.50	3.72	3.67	4.04	4.74	4.99	4.64	4.76	4.63	3.07	3.84	5.00	4.04	4.12
	pt	2.71	4.10	5.00	3.65	4.00	3.89	4.16	4.97	5.00	4.87	4.96	4.79	3.26	4.25	5.00	4.18	4.40
Medium-Resource	pl	2.37	3.04	5.00	3.19	3.25	3.37	4.17	4.37	5.00	4.24	4.40	4.44	3.24	3.79	5.00	3.97	4.06
	tr	2.26	2.64	5.00	2.86	2.91	3.13	3.59	3.97	5.00	3.72	3.99	4.05	2.60	2.35	5.00	2.60	2.65
	avg.	2.53	3.73	5.00	3.47	3.70	3.69	3.99	4.65	5.00	4.50	4.66	4.56	3.13	3.95	5.00	3.98	4.09
	vi	2.21	3.98	5.00	3.46	3.78	3.69	4.21	4.83	5.00	4.59	4.82	4.69	2.89	3.46	5.00	3.47	3.61
	id	2.83	3.97	5.00	3.70	3.90	3.88	4.21	4.91	5.00	4.68	4.84	4.73	3.38	3.80	5.00	3.92	3.96
	ko	2.45	3.72	5.00	3.30	3.46	3.59	3.31	4.05	5.00	3.76	3.88	4.00	2.88	3.69	4.99	3.64	3.73
	sv	2.94	3.29	5.00	3.51	3.69	3.69	3.97	4.85	5.00	4.52	4.76	4.62	3.35	4.03	5.00	4.10	4.24
	ar	2.23	2.66	5.00	2.69	2.75	3.07	3.90	4.10	5.00	3.99	4.09	4.22	2.71	2.77	5.00	2.85	2.93
	hu	2.24	2.47	5.00	2.61	2.84	3.03	4.00	4.32	5.00	4.19	4.35	4.37	3.14	3.49	5.00	3.69	3.73
	el	1.98	1.60	5.00	1.91	1.99	2.50	3.73	3.63	5.00	3.69	3.91	3.99	2.47	1.98	5.00	2.25	2.25
Low-Res.	uk	2.59	3.52	5.00	3.39	3.61	3.62	4.38	4.68	5.00	4.64	4.70	4.68	3.51	4.30	5.00	4.30	4.43
	da	3.07	3.76	5.00	3.61	3.88	3.86	4.12	4.67	5.00	4.56	4.72	4.61	3.40	3.99	5.00	4.12	4.21
	th	2.54	3.46	5.00	3.03	3.29	3.46	3.63	4.35	5.00	4.02	4.26	4.25	2.38	2.50	5.00	2.53	2.55
	fi	2.27	2.30	5.00	2.67	2.80	3.01	3.44	3.36	4.98	3.46	3.78	3.80	2.21	1.93	5.00	2.18	2.28
Low-Res.	hr	2.57	2.94	5.00	3.11	3.24	3.37	4.16	3.67	5.00	3.88	4.12	4.17	3.58	3.51	5.00	3.88	3.95
	hi	2.29	3.19	5.00	3.01	3.13	3.32	3.73	4.72	4.98	4.37	4.60	4.48	2.76	2.79	5.00	2.83	2.88
	bn	2.36	2.61	5.00	2.45	2.53	2.99	3.80	4.01	4.98	3.69	3.94	4.08	2.13	2.00	5.00	2.00	1.98
	avg.	2.47	3.10	5.00	3.03	3.21	3.36	3.90	4.30	5.00	4.15	4.34	4.34	2.91	3.16	5.00	3.27	3.34

Table 21: Detailed Common Grounds Evaluation with GPT4o-as-a-judge. Average over the evaluated Common grounds for each model and language is reported. In bold, is the best rating among the models for each criterion.

Lang.	Eval. Source	Models																
		Gemma-1.1-7b								LLaMA3.1-8B								
		S	F	Tx	Relevance		Avg.	S	F	Tx	Relevance		Avg.	S	F	Tx	Relevance	
es	GPT4o	2.75	4.17	5.00	3.58	4.00	3.90	4.00	5.00	5.00	5.00	5.00	4.80	3.40	4.60	5.00	4.47	4.73
	Human	3.25	1.75	5.00	3.17	3.75	3.38	4.71	4.00	5.00	3.86	4.00	4.31	3.87	2.80	5.00	3.67	4.33
zh	GPT4o	1.86	3.86	5.00	3.29	3.71	3.54	3.88	4.50	5.00	4.25	4.63	4.45	2.70	4.00	5.00	3.70	3.80
	Human	4.14	4.29	4.86	3.71	4.14	4.23	5.00	5.00	5.00	4.88	5.00	4.98	4.50	4.20	4.70	4.60	4.50
fr	GPT4o	2.53	4.00	5.00	3.47	3.65	3.73	4.29	5.00	5.00	4.82	5.00	4.82	3.56	4.25	5.00	4.19	4.38
	Human	5.00	5.00	4.94	4.29	4.53	4.75	5.00	4.65	5.00	4.65	4.82	4.82	4.69	4.50	4.81	4.12	4.52
vi	GPT4o	1.89	4.00	5.00	3.56	3.78	3.64	4.14	4.57	5.00	4.29	4.71	4.54	2.71	3.29	5.00	3.43	3.57
	Human	3.33	4.44	5.00	4.22	4.22	4.24	4.43	5.00	5.00	4.57	4.86	4.77	3.00	3.14	4.29	3.29	3.57
ar	GPT4o	1.97	2.65	5.00	2.55	2.61	2.95	3.92	4.06	5.00	4.03	4.08	4.22	2.90	2.80	5.00	3.10	3.15
	Human	3.06	3.29	4.77	3.55	3.65	3.66	4.39	4.14	4.97	4.19	4.42	4.42	3.60	3.00	4.60	3.75	3.60

Table 22: Detailed Human Evaluations of Common Grounds and Comparison with LLM judgments on the Same Data Points. Average over the evaluated common grounds for each model and language is reported. In bold, is the best rating among the models for each criterion.

Lang.		Models																								
		Gemma-1.1-7b								LLaMA3.1-8B								Mistral-7B								
		S	F	H	Tx	Relevance		P	CGT	Avg.	S	F	H	Tx	Relevance		P	CGT	Avg.	S	F	H	Tx	Relevance		P
High-Resource	en	2.86	4.73	2.56	5.00	4.30	4.64	4.01		3.83	4.97	3.99	4.97	4.97	4.94	4.61	3.30	4.90	3.26	5.00	4.86	4.82	4.36			
	ru	1.89	3.11	1.89	5.00	3.07	3.24	3.03		3.90	4.77	3.61	4.98	4.80	4.80	4.48	3.12	4.15	2.52	5.00	4.33	4.33	3.91			
	de	2.36	3.52	2.05	5.00	3.49	3.72	3.36		3.98	4.72	3.71	5.00	4.73	4.78	4.49	2.88	3.75	2.35	5.00	4.29	4.25	3.75			
	jp	2.06	3.31	1.91	5.00	3.04	3.09	3.07		3.49	4.19	3.18	5.00	4.06	4.11	4.01	2.57	3.03	2.06	5.00	3.15	3.22	3.17			
	es	2.47	4.26	2.14	5.00	3.86	4.07	3.63		4.10	5.00	4.01	5.00	4.96	4.97	4.67	3.28	4.19	2.53	5.00	4.37	4.39	3.96			
	zh	2.08	3.83	1.98	5.00	3.44	3.53	3.31		3.75	4.74	3.68	5.00	4.59	4.62	4.40	2.79	4.10	2.49	5.00	4.10	4.12	3.77			
	fr	2.59	3.89	2.08	5.00	3.50	3.68	3.46		4.27	4.93	3.85	5.00	4.92	4.95	4.63	3.31	4.29	2.49	5.00	4.32	4.43	3.97			
	it	3.06	3.69	2.06	5.00	3.62	3.78	3.54		4.59	4.94	4.05	4.98	4.83	4.90	4.72	3.42	3.94	2.60	5.00	4.13	4.22	3.88			
	nl	2.47	3.24	2.10	5.00	3.45	3.58	3.31		4.06	4.79	3.81	4.98	4.78	4.82	4.54	3.02	3.68	2.46	5.00	4.14	4.17	3.74			
	pt	2.59	3.92	2.10	5.00	3.61	3.84	3.51		4.18	4.97	3.95	5.00	4.94	4.95	4.67	3.21	4.32	2.44	5.00	4.41	4.46	3.97			
Medium-Resource	pl	2.21	2.75	1.81	5.00	3.13	3.17	3.01		4.19	4.38	3.44	5.00	4.38	4.48	4.31	3.13	3.38	2.30	5.00	3.99	3.98	3.63			
	tr	2.13	2.40	1.67	5.00	2.74	2.76	2.78		3.59	4.02	3.03	5.00	3.92	4.11	3.95	2.45	2.08	1.65	5.00	2.40	2.47	2.68			
	avg.	2.40	3.55	2.03	5.00	3.44	3.59	3.34		3.99	4.70	3.69	4.99	4.66	4.70	4.46	3.04	3.82	2.43	5.00	4.04	4.07	3.73			
	vi	2.09	3.80	2.10	5.00	3.44	3.65	3.35		4.20	4.92	3.96	5.00	4.87	4.86	4.64	2.79	3.29	2.24	5.00	3.48	3.50	3.38			
	id	2.68	3.77	2.31	5.00	3.69	3.81	3.54		4.23	4.94	4.07	5.00	4.83	4.85	4.65	3.33	3.69	2.38	5.00	3.89	3.86	3.69			
	ko	2.26	3.44	1.97	5.00	3.25	3.30	3.20		3.28	4.00	2.90	5.00	3.78	3.90	3.81	2.77	3.47	2.04	5.00	3.58	3.59	3.41			
	sv	2.68	3.13	2.07	5.00	3.45	3.64	3.33		4.02	4.85	3.73	4.98	4.72	4.80	4.52	3.32	3.77	2.35	5.00	4.28	4.32	3.84			
	ar	2.04	2.29	1.61	5.00	2.55	2.67	2.69		3.91	4.12	3.16	4.99	4.23	4.19	4.10	2.61	2.56	1.77	5.00	2.76	2.83	2.92			
	hu	2.07	2.21	1.56	5.00	2.50	2.75	2.68		4.00	4.37	3.31	5.00	4.37	4.44	4.25	3.02	3.18	2.13	5.00	3.69	3.70	3.45			
	el	1.61	1.20	1.15	5.00	1.67	1.89	2.09		3.70	3.53	2.98	5.00	3.70	3.87	3.80	2.24	1.61	1.25	5.00	2.09	2.18	2.40			
Low-Res.	uk	2.44	3.18	1.99	5.00	3.33	3.42	3.23		4.39	4.66	3.79	5.00	4.70	4.73	4.55	3.48	4.07	2.74	5.00	4.36	4.39	4.01			
	da	2.90	3.50	2.26	5.00	3.59	3.78	3.51		4.12	4.67	3.70	5.00	4.76	4.77	4.50	3.38	3.70	2.45	5.00	4.16	4.13	3.80			
	th	2.21	3.09	1.82	5.00	2.84	3.07	3.01		3.60	4.43	3.06	5.00	4.23	4.37	4.12	2.11	2.09	1.52	5.00	2.31	2.33	2.56			
	fi	1.93	2.02	1.44	5.00	2.55	2.69	2.60		3.44	3.34	2.58	4.98	3.49	3.83	3.61	2.08	1.72	1.40	5.00	2.18	2.27	2.44			
	hr	2.31	2.54	1.74	5.00	2.90	3.02	2.92		4.16	3.63	3.02	4.99	3.96	4.15	3.99	3.50	3.12	2.28	5.00	3.84	3.88	3.60			
	hi	2.23	2.93	1.95	5.00	2.86	2.94	2.98		3.74	4.74	3.80	4.97	4.65	4.73	4.44	2.51	2.52	1.68	5.00	2.64	2.67	2.84			
	bn	2.18	2.27	1.62	5.00	2.33	2.48	2.65		3.80	4.00	4.98	3.79	3.96	3.92	3.89	1.89	1.65	1.27	5.00	1.83	1.82	2.24			
	avg.	2.26	2.81	1.83	5.00	2.92	3.08	2.98		3.90	4.30	3.36	4.99	4.29	4.39	4.21	2.79	2.89	1.96	5.00	3.22	3.25	3.18			
Human	af	2.51	3.28	2.10	5.00	3.16	3.43	3.25		3.92	4.25	3.38	5.00	4.40	4.51	4.24	2.82	2.34	1.78	5.00	3.12	3.17	3.04			
	sw	1.44	1.70	1.18	5.00	1.69	1.75	2.13		3.35	3.53	2.66	5.00	3.37	3.49	3.57	1.32	1.23	1.04	5.00	1.22	1.22	1.84			
	yo	2.22	1.95	1.39	4.94	1.91	2.38	2.35	1.66	1.25	5.00	1.64	2.26	2.14	1.51	1.15	5.00	1.57	1.57	2.16						
	avg.	2.06	2.31	1.56	4.98	2.24	2.36	2.58		3.21	3.15	2.43	5.00	3.14	3.21	3.36	2.09	1.69	1.32	5.00	1.97	1.99	2.34			
GPT4o	es	2.50	4.08	2.08	5.00	3.58	3.92	3.53		4.00	5.00	4.00	5.00	5.00	5.00	4.67	3.40	4.47	2.53	5.00	4.67	4.73	4.13			
	zh	1.71	3.86	2.00	5.00	3.43	3.71	3.29		3.88	4.63	3.63	5.00	4.63	4.75	4.42	2.60	3.90	2.20	5.00	3.70	3.80	3.53			
	fr	2.53	4.00	2.18	5.00	3.47	3.53	3.45		4.29	5.00	4.00	5.00	5.00	5.00	4.72	3.50	4.25	2.56	5.00	4.25	4.38	3.99			
	vi	1.67	4.00	2.00	5.00	3.56	3.67	3.31		4.14	4.86	4.14	5.00	5.00	4.86	4.67	2.71	3.14	2.00	5.00	3.43	3.29	3.26			
	ar	1.87	2.48	1.52	5.00	2.48	2.58	2.66		3.94	4.00	3.11	5.00	4.22	4.19	4.08	2.75	2.50	1.75	5.00	2.95	3.05	3.00			
	avg.	2.77	2.55	1.87	4.74	3.03	3.35	3.05		4.25	3.58	3.47	4.83	4.28	4.39	4.13	3.70	2.80	2.45	4.40	3.70	3.80	3.48			

Table 23: Detailed Conversations Evaluation with GPT4o-as-a-judge. Average over the evaluated conversations for each model and language is reported. In bold, is the best rating among the models for each criterion.

Table 24: Detailed Human Evaluations of Conversations and Comparison with LLM judgments on the Same Data Points. Average over the evaluated conversations for each model and language is reported. In bold, is the best rating among the models for each criterion.

G Evaluation Platform

If you don't have an ID, please fill the following form to get one:

Language:	Choose your language			
Age:	Under 18: <input type="radio"/>	18 - 29: <input type="radio"/>	30 - 49: <input type="radio"/>	50 +: <input type="radio"/>
Gender:	Female: <input type="radio"/>	Male: <input type="radio"/>	Other: <input type="radio"/>	Prefer Not To Say: <input type="radio"/>
Education level:	High-school or less: <input type="radio"/>	University degree (Undergraduate, Bachelor, Master): <input type="radio"/>	Ph.D. or higher: <input type="radio"/>	
Work status:	Student: <input type="radio"/>	Out-of-work: <input type="radio"/>	Employed: <input type="radio"/>	
How did you find us ?	Social network: <input type="radio"/>	Word of mouth/Referral: <input type="radio"/>	Mailing list: <input type="radio"/>	Direct contact from authors: <input type="radio"/>

Figure 3: Demographic Form Completed by Users at their First Login on the Evaluation Platform

Definitions and evaluation's guidelines

As an evaluator, your task is to carefully assess the following components. For each component, several assertions are given. You may evaluate a component by giving your opinion about each assertion.

Personas:
You will be presented with two personas, each described by five profile sentences. These sentences follow a specific taxonomy and define the character's background, traits, and perspective. Evaluate whether the personas are consistent with the taxonomy and clearly distinguishable in the conversation.

Common Ground:
The Common Ground refers to the shared context and joint activity that serves as the foundation for the conversation, much like how real-life discussions are anchored in specific situations. Ensure that the conversation aligns with this context and remains coherent in relation to the personas and the keywords provided as common ground taxonomy (e.g., gossip, serious talk, etc.).

Conversation:
The dialogue between the two speakers should reflect a back-and-forth interaction rooted in the Common Ground. Both speakers must stay true to their personas throughout the exchange. Assess whether the conversation maintains this coherence, staying logical and true to the taxonomy, personas, and context.

Your role is critical in helping us ensure that each aspect of the conversation is aligned and contributes to a realistic and meaningful interaction. Please take your time to carefully evaluate each element based on the provided criteria. Each of which is assessed on a **1 to 5 scale**, with 1 being the worst rating and 5 the best.
Thank you for your participation!

Figure 4: Additional Guidelines Before Each Conversation's Evaluation on the Platform

Persona 1

Profiles	Taxonomies			
Evaluation				
Specificity: The persona is specific to the language, incorporating relevant entities such as names, cities, culture, and folk psychology when possible or at least does not include elements from another language or culture.				
1: Strongly disagree <input type="radio"/>	2: Disagree <input type="radio"/>	3: Neutral <input type="radio"/>	4: Agree <input type="radio"/>	5: Strongly agree <input type="radio"/>
Fluency: The language quality of the provided persona is good, the text is plausible, and the provider demonstrates strong language skills.				
1: Strongly disagree <input type="radio"/>	2: Disagree <input type="radio"/>	3: Neutral <input type="radio"/>	4: Agree <input type="radio"/>	5: Strongly agree <input type="radio"/>
Relevancy: Each persona's sentences are relevant to the provided taxonomy, and all sentences within the persona are coherent with one another, with no contradictory facts.				
1: Strongly disagree <input type="radio"/>	2: Disagree <input type="radio"/>	3: Neutral <input type="radio"/>	4: Agree <input type="radio"/>	5: Strongly agree <input type="radio"/>
Toxicity: The persona is free from toxicity and does not contain harmful or offensive content.				
1: Strongly disagree <input type="radio"/>	2: Disagree <input type="radio"/>	3: Neutral <input type="radio"/>	4: Agree <input type="radio"/>	5: Strongly agree <input type="radio"/>

Figure 5: Persona's Human Evaluation From

Conversation content

Evaluation				
Specificity: The conversation is specific to the language, incorporating relevant entities such as names, cities, culture, and folk psychology, or at least does not include elements from another language or culture.				
1: Strongly disagree <input type="radio"/>	2: Disagree <input type="radio"/>	3: Neutral <input type="radio"/>	4: Agree <input type="radio"/>	5: Strongly agree <input type="radio"/>
Humanness: The conversation feels more natural and authentic rather than automated or artificial.				
1: Strongly disagree <input type="radio"/>	2: Disagree <input type="radio"/>	3: Neutral <input type="radio"/>	4: Agree <input type="radio"/>	5: Strongly agree <input type="radio"/>
Fluency: The language quality of the speakers is good, the messages are plausible and strong language skills are demonstrated.				
1: Strongly disagree <input type="radio"/>	2: Disagree <input type="radio"/>	3: Neutral <input type="radio"/>	4: Agree <input type="radio"/>	5: Strongly agree <input type="radio"/>
Toxicity: The conversation is free from toxicity and does not contain harmful or offensive content.				
1: Strongly disagree <input type="radio"/>	2: Disagree <input type="radio"/>	3: Neutral <input type="radio"/>	4: Agree <input type="radio"/>	5: Strongly agree <input type="radio"/>
Relevancy to personas: The conversation is consistent and faithful to both speakers' personas, without contradictory elements.				
1: Strongly disagree <input type="radio"/>	2: Disagree <input type="radio"/>	3: Neutral <input type="radio"/>	4: Agree <input type="radio"/>	5: Strongly agree <input type="radio"/>
Relevancy to common ground: The conversation is consistent and faithful to the common-ground context provided, and clearly reflects the associated taxonomy type.				
1: Strongly disagree <input type="radio"/>	2: Disagree <input type="radio"/>	3: Neutral <input type="radio"/>	4: Agree <input type="radio"/>	5: Strongly agree <input type="radio"/>

Figure 6: Conversation's Human Evaluation From

H Examples of Conversations from MOUD

P1's Persona	Taxonomy
I'm a mediocre guitarist, but I love playing acoustic. I'm unhappy with my job as a data entry clerk. I moved to portland, oregon, for the food scene. I save money by cooking at home every night. My bike is a 2015 trek mountain bike.	Psychographics Personal Characteristics Personality Trait Creativity Demographics Employment Job fulfillment Demographics Location Residence Psychographics Personal Characteristics Financial Awareness Budgeting Demographics Possession Vehicle
P2's Persona	Taxonomy
I follow a vegan diet for health reasons. I practice yoga to reduce stress and anxiety. I dislike martin scorsese's films, find them too dark. I hold a degree in environmental science. I'm currently studying at the university of california.	Wellness Diet Dietary Restrictions Wellness Preventive Care Alternative medication Psychographics Preference Movie Director Demographics Education Degree Demographics Education School name

Common Ground: Goal-directed talk | Interrogation

Character 1, a mediocre guitarist, meets Character 2, a vegan environmental science student, at a Portland farmers market, where they both frequent to buy fresh produce. Character 1 is intrigued by Character 2's extensive knowledge of sustainable living and decides to grill them with questions, hoping to gain insights on how to live a more environmentally conscious lifestyle while also pursuing his passion for music. Character 2, aware of Character 1's dissatisfaction with his job, is happy to engage in a conversation that combines their shared interest in sustainability and music.

Conversation

P1: What inspired you to study environmental science, and how does it relate to your daily life?
 P2: Growing up, I was always concerned about the planet's future, so I chose environmental science to make a difference.
 P1: That's admirable, how do you think I, as a musician, can contribute to sustainability efforts in Portland?
 P2: As a musician, you can raise awareness through your performances and songs, inspiring others to adopt eco-friendly habits.
 P1: That's a great point, do you think music festivals like Pickathon could do more to reduce their environmental impact?
 P2: Yes, they could implement more sustainable practices, like composting and reducing single-use plastics, to minimize their carbon footprint.
 P1: Composting and reducing waste would be a huge step, have you noticed any changes in the local food scene since you moved here?
 P2: Yes, Portland's vegan options have exploded, which aligns with my dietary choices, and I appreciate the emphasis on locally sourced ingredients.
 P1: Locally sourced is a big deal to me too, do you know any good spots to grab a vegan meal on a budget in the city?
 P2: I love PSweetpea Baking Co. for affordable, plant-based bowls and pastries, and they're committed to sustainability too.

Table 25: English Conversation from MOUD with Speech Event Type: Goal-Directed talk | Interrogation

P1's Persona	Taxonomy
I've participated in beach cleanups every summer since I was a kid. I avoid watching reality TV shows. My go-to browser is brave for its security features. I'm a close friend to my childhood best friend. I currently live in portland, oregon.	Psychographics Interests Environment Psychographics Preference Media Genre Demographics Possession Tech Device Psychographics Personal Characteristics Social Connections Demographics Location Residence
P2's Persona	Taxonomy
I have a recurring knee pain from playing basketball. I drink a glass of wine to unwind after work. My mom has diabetes and has to take medication. I get a physical every 6 months due to my job. I wake up at 5 am every day.	Wellness Symptom Physical Symptom Wellness Substance Use Alcohol Consumption Wellness Medical History Wellness Preventive Care Check-ups Wellness Sleep Sleep Quality

Common Ground: Informal/Superficial talk | Reminiscing

It's a crisp summer evening in Portland, Oregon, and Character 1 and Character 2 are sitting on a patio overlooking the Willamette River, reminiscing about their past adventures. They're catching up after a while, and Character 1 mentions they're planning to organize a beach cleanup event soon. Character 2, who has been dealing with recurring knee pain from basketball, is happy to join in, but first wants to share a funny story about their experience volunteering at a beach cleanup together. Character 2 starts telling Character 1 about the time they accidentally got stuck in a trash can during a previous cleanup event, making Character 1 laugh out loud.

Conversation

P1: I'm glad you're feeling up to joining me for the beach cleanup, my friend!
 P2: I'm in, but I've got to laugh first - remember that time we got stuck in that trash can?
 P1: Oh man, how could I forget? You were wedged in like a sardine!
 P2: I swear, I'm still laughing about it, my knee was throbbing from basketball, but that was a worse pain!
 P1: And I'm pretty sure I almost got a hernia trying to pull you out!
 P2: I'm just glad I didn't have to miss work the next day, my physical was the week after that!
 P1: You got lucky, my dad was still friends with your dad, so I'm sure he'd have covered for you somehow!
 P2: Yeah, my dad always joked that your dad was the only one who could get him out of trouble!

Table 26: English Conversation from MOUD with Speech Event Type: Informal/Superficial talk | Reminiscing

For both Table 25 and Table 26 which represent English conversations, colored texts highlight how Common Grounds and Conversations incorporate elements related to the Persona 1, the Persona 2, and the type of Speech Event.

In Table 25, we see how the SE (Goal-Directed talk | Interrogation) is introduced in the CG with "decides to grill them with questions, hoping to gain insights on how to live a more environmentally conscious lifestyle" and materialized in the conversation by the questions at each turn from Persona 1.

In the meantime, there are multiple references to environment, vegan lifestyle, music or Portland all related to the personas involved. Apart from the personas and SE elements incorporated to the CG and the conversation, we observe references to elements specific to Oregon, USA, as one character's persona mentions they moved to Portland: Sweetpea Baking Co., Pickathon, etc. These details showcase the cultural specificity we aimed for in the dataset. It may seem obvious in English examples. For a clearer understanding of why it is not and why MT is limiting in preserving cultural nuance, see in Table 27 which features a French conversation.

In Table 26, we have a completely different type of SE: Informal/Superficial talk | Reminiscing. Here the speakers' roles are symmetric, a story connecting both speakers from the past and based on their personas in the CG is created . The conversations, incorporate it along with speakers personas.

P1's Persona	Taxonomy
Je préfère utiliser un macbook pour travailler. (<i>I prefer to use a macbook for work.</i>) Je suis freelance, ce qui me permet de travailler à domicile. (<i>I'm a freelancer, which allows me to work from home.</i>)	Psychographics Preferences Favorite Apps Demographics Employment Job fulfillment
Mon film préféré est Les intouchables. (<i>My favorite film is Les intouchables.</i>) Je vis dans une maison de campagne avec mon chien. (<i>I live in a country house with my dog.</i>) Je suis en train d'apprendre le japonais pour voyager. (<i>I'm learning Japanese to travel.</i>)	Psychographics Preference Movie Title Demographics Socioeconomic Status Housing status Demographics Education Spoken Languages
P2's Persona	Taxonomy
Je suis d'origine corse, j'adore la mer méditerranée. (<i>I'm native Corsican and I love the Mediterranean Sea.</i>) Je suis fan de la chanson "La Vie en rose" de Édith Piaf. (<i>I'm a fan of the song "La Vie en rose" by Edith Piaf.</i>) Je voudrais un ordinateur portable portable. (<i>I would like a laptop computer.</i>) Je préfère lire "Les Misérables". (<i>I enjoy reading "Les Misérables".</i>) Je fais de la randonnée dans les Alpes. (<i>I go hiking in the Alps.</i>)	Demographics Employment Company name Psychographics Influences Role Models Demographics Possession Tech Device Psychographics Preference Book Title Wellness Fitness Exercise Routine
Common Ground : Goal-directed talk Asking out	
Personnage 1 et Personnage 2 se rencontrent lors d'un festival de musique à la plage, dans un charmant village côtier corse. Personnage 1, qui a été impressionné par la voix de Personnage 2 lors d'une performance de chant de la chanson "La Vie en rose" d'Édith Piaf, décide de l'aborder pour lui proposer de partager un pique-nique sur la plage la semaine prochaine, profitant du coucher de soleil et de la vue imprenable sur la mer Méditerranée. (Character 1 and Character 2 meet at a beach music festival in a charming Corsican coastal village. Character 1, who was impressed by Character 2's voice during a singing performance of Édith Piaf's song "La Vie en rose", decides to approach her to propose sharing a picnic on the beach next week, enjoying the sunset and the breathtaking view over the Mediterranean Sea.)	
Conversation	
P1: Vous avez chanté magnifiquement "La Vie en rose" aujourd'hui, vous êtes vraiment doué(e) ! (<i>You sang "La Vie en rose" beautifully today, you're really talented!</i>)	
P2: Merci beaucoup, c'est un grand honneur d'avoir été apprécié ! (<i>Thank you very much, it's a great honor to have been appreciated!</i>)	
P1: Je me demandais si vous aimeriez partager un pique-nique avec moi sur la plage la semaine prochaine, pour admirer le coucher de soleil ? (<i>I was wondering if you'd like to share a picnic with me on the beach next week, to watch the sunset?</i>)	
P2: Quel joli plan ! Je me sens déjà en vacances. Quelle plage préférez-vous ? (<i>What a nice plan! I feel like I'm already on vacation. Which beach do you enjoy most?</i>)	
P1: J'ai pensé à la plage de Sainte-Maxime, elle est magnifique à ce moment-là de l'année, le ciel se teinte de couleurs magnifiques. (<i>I thought of the beach at Sainte-Maxime, which is magnificent at this time of year, the sky tinted with magnificent colors.</i>)	
P2: C'est l'une de mes préférées ! Je connais bien cette plage, c'est idéal pour regarder le coucher de soleil. Quand avions-nous l'idée de nous y rendre ? (<i>It's one of my favorites! I know this beach well, and it's ideal for watching the sunset. When do we have the idea of going there?</i>)	
P1: Je pensais que nous pourrions y aller le samedi prochain, je me libère de mon travail à 17h, nous pourrions partir à 18h et profiter du coucher de soleil ensemble. (<i>I was thinking we could go next Saturday, I get off work at 5 pm, we could leave at 6 pm and enjoy the sunset together.</i>)	
P2: Je suis libre également ce jour-là, mais je devrai ramener mon ordinateur portable pour travailler un peu, tu es d'accord ? (<i>I'm free that day too, but I'd like to take my laptop with me and do a bit of work, is that okay with you?</i>)	
P1: Pas de problème, tu peux apporter ton MacBook, je comprends que les freelances ont besoin de rester productifs même en vacances ! (<i>No worries, you can bring your MacBook, I understand that freelancers need to stay productive even on vacation!</i>)	
P2: Merci de comprendre ! Et si on dîne ensemble après le coucher de soleil, peut-être dans un petit resto à côté de la plage ? (<i>Thanks for your understanding! How about dinner after sunset, maybe at a small restaurant by the beach?</i>)	
P1: Excellente idée, j'ai déjà réservé une table au Bistrot de la Plage, leur cuisine est délicieuse et la vue sur la mer est splendide. (<i>Excellent idea, I've already booked a table at the Bistrot de la Plage, their food is delicious and the sea view is splendid.</i>)	
P2: C'est parfait ! Je serais ravi de partager un dîner romantique avec toi, à 20h ? (<i>It's fantastic! I'd be delighted to share a romantic dinner with you, say at 8 pm?</i>)	
P1: Je serais ravi de t'avoir à côté de moi, je t'attendrai à 20h au Bistrot de la Plage, avec un bouquet de fleurs pour toi ! (<i>I'd be delighted to have you next to me, and I'll be waiting for you at 8pm at the Bistrot de la Plage, with a bouquet of flowers for you!</i>)	
P2: Je suis ému, merci pour l'attention ! Je te rejoindrai avec un grand sourire et un cœur léger, en espérant que la soirée sera magique ! (<i>I'm so touched, thank you for your attention! I'll join you with a big smile and a light heart, hoping for a magical evening!</i>)	

Table 27: Example of French Conversation with *translations* and highlights on notable aspects of MOUD: in orange are elements Specific to French Culture and Geography and in red those related to Type of Speech Event at hand which here is Asking Out.

The example in Table 27 is a French conversation to emphasizes on another key improvement of MOUD over existing similar multilingual datasets (often based on MT) in languages different from English. Indeed, are highlighted elements specific to French Culture, such as "*Les Misérables*", "*Les Intouchables*", "*Edith Piaf*", "*Corsica*", or the "*Saint-Maxime Beach*". These are details that would not have been generated natively with MT, leading to French conversations that lacked cultural specificity. Furthermore, we again observe elements related to the type of SE (Goal-directed talk | Asking out) in the CG and the Conversation. One can notice how the flow of the chat is different from that of the previous examples. The first character is trying to seduce the second and invite them to go out next week which pleases the character 2.

I Speech events Taxonomy

Highlighted in blue, are elements added to taxonomy to enhance the *understanding* of the LLM, to promote diversity through reformulations, and each Speech Event speakers' roles symmetry to facilitate the creation of adequate dialogue.

I.1 INVOLVING TALK AUGMENTED

Cat	Sub Category	Description	Reformulations	S1=S2
Involving Talk	Making up	Speaker 1 apologizes to Speaker 2 or both apologize for violating some expectations.	Speaker 1 is apologizing to Speaker 2. S1 is making up with S2 after a disagreement. S1 & S2 are mending their relationship.	✓
	Love talk	The speakers are expressing love and giving attention and affection.	S1 & S2 are sharing affectionate words. S1 & S2 are expressing their love for each other. S1 & S2 are engaging in a loving conversation.	✓
	Relationship talk	The speakers are talking about the nature and state of their relationship.	S1 & S2 are in a heated discussion. S1 & S2 having a disagreement. S1 & S2 having a conflict in their conversation.	✓
	Serious conversation	The speakers are having an in-depth discussion or exchange of feelings, opinions, or ideas about a personal and important topic.	S1 & S2 are joking around for fun. S1 & S2 are telling jokes to lighten the mood. S1 & S2 are engaging in playful banter.	✓
	Talking about problems	G: S1 is telling about some problem, while S2 is trying to help. S1: The speaker is breaking bad news to their interlocutor. S2: The speaker is receiving bad news from their interlocutor.	S1 is explaining a problem to S2. S2 is offering help to S1 for a problem. S1 is seeking advice from Speaker S2.	✗
	Breaking bad news	G: S1 is telling some bad news that S2 doesn't know about. S1: The speaker is breaking bad news to their interlocutor. S2: The speaker is receiving bad news from their interlocutor.	S1 is informing S2 about something unfortunate. S1 is revealing bad news to S2. S1 is telling S2 something they didn't want to hear.	✗
	Complaining	The speakers are expressing negative feelings, frustrations, gripes, or complaints toward some common experience.	S1 & S2 are expressing their dissatisfaction. S1 & S2 complaining about a shared issue. S1 & S2 venting their frustrations.	✓

Table 28: Taxonomy of Speech Events of the category Involving Talk.

I.2 GOAL-DIRECTED TALK AUGMENTED

Cat	Sub Category	Description	Reformulations	S1=S2
Goal-directed Talk	Persuading conversation	G: S1 is convincing S2 to do something. S1: The speaker is trying to convince their interlocutor to do something. S2: The speaker is being persuaded by their interlocutor to take action.	S1 is convincing S2 to agree to something. S1 is trying to persuade S2 to take action. S1 is attempting to sway S2's opinion.	✗
	Decision-making conversation	The speakers are working towards making a decision about some task.	S1 & S2 are deciding on a course of action. S1 & S2 are discussing their options. S1 & S2 are weighing the pros and cons.	✓
	Giving and getting instructions	G: S1 is giving S2 information or directions about how to do some task. S1: The speaker is giving instructions to their interlocutor on how to do something. S2: The speaker is receiving instructions from their interlocutor.	S1 is instructing S2 on how to complete a task. S1 is giving directions to S2. S1 is telling S2 the steps to follow.	✗
	Class information Talk	The speakers are having informal conversations to find out about class assignments, exams, or course material.	S1 & S2 are discussing class assignments. S1 & S2 are exchanging information about their classes. S1 & S2 are reviewing course-related topics.	✓
	Lecture	G: S1 is telling S2 how to act or what to do in a one-way conversation. S1: The speaker is telling their interlocutor how to act or what to do. S2: The speaker is listening to instructions or advice from their interlocutor.	S1 providing guidance S2 without expecting a response. S1 is lecturing S2 on how to behave. S1 is telling S2 what they should do.	✗
	Interrogation	G: S1 is grilling S2 with questions. S1: The speaker is asking probing questions to their interlocutor. S2: The speaker is responding to the probing questions from their interlocutor.	S1 providing guidance S2 without expecting a response. S1 is lecturing S2 on how to behave. S1 is telling S2 what they should do.	✗
	Making Plans	The speakers are talking to arrange a meeting or to do something together.	S1 & S2 are arranging a time to get together. S1 & S2 are discussing what to do together. S1 & S2 are coordinating their schedules.	✓
Asking a favor		G: S1 is getting S2 to do something for them. S1: The speaker is asking their interlocutor for a favor. S2: The speaker is considering whether to grant the favor requested by their interlocutor.	S1 is requesting help from S2. S1 trying to get S2 to do something for them. S1 trying to get S2 to agree to a favor.	✗
	Asking out	G: S1 is asking S2 out on a date. S1: The speaker is asking their interlocutor out on a date. S2: The speaker is considering whether to go on a date with their interlocutor.	S1 is asking S2 to go on a date. S1 is inviting S2 out. S1 is asking S2 to spend time together romantically.	✗

Table 29: Taxonomy of Speech Events of the category Goal-directed Talk.

I.3 INFORMAL / SUPERFICIAL TALK AUGMENTED

The underlined subcategory "*Getting to know someone*" is the most common in personas-based ODD datasets hence restricting their actual openness. In this category's speech event types the speaker always have equivalent roles in the conversation's flow.

Cat	Sub Category	Description	Reformulations	S1=S2
Informal/Superficial Talk	Small Talk	The speakers are passing time and avoiding being rude.	Speaker 1 and Speaker 2 are making small talk to pass time. S1 & S2 are talking casually to be polite. S1 & S2 are chatting to avoid awkward silence.	✓
	Currents events talk	The speakers are talking about news and current events.	S1 & S2 are discussing today's top stories. S1 & S2 are sharing opinions on the latest headlines. S1 & S2 are conversing about what's happening in the world.	✓
	Gossip	The speakers are exchanging opinions or information about someone else when that person isn't present.	S1 & S2 are sharing rumors about another person. S1 & S2 are discussing someone else's business. S1 & S2 talking about someone behind their back.	✓
	Joking around	The speakers are engaging in a playful kind of talk to have fun or release tension.	S1 & S2 are joking around for fun. S1 & S2 are telling jokes to lighten the mood. S1 & S2 are engaging in playful banter.	✓
	Catching up	The speakers are talking about the events that have occurred since they last spoke.	S1 & S2 are updating each other on their lives. S1 & S2 are talking about what's been happening. S1 & S2 are sharing what's new since they last spoke.	✓
	Recapping the day's events	The speakers are telling each other about what's happened to them during the day.	S1 & S2 are talking about the highlights of their day. S1 & S2 are discussing how their day went. S1 & S2 are recounting the day's experiences.	✓
	<u>Getting to know someone</u>	The speakers are getting acquainted with each other.	S1 & S2 are discussing today's top stories. S1 & S2 are sharing opinions on the latest headlines S1 & S2 are conversing about what's happening in the world.	✓
	Sports talk	The speakers are talking about playing or watching a sporting event.	S1 & S2 are discussing a sporting event. S1 & S2 are analyzing the performance of a sports team. S1 & S2 are debating the outcome of a recent game.	✓
	Morning talk	The speakers are engaging in routine talk when waking up in the morning.	S1 & S2 are discussing their plans for the day. S1 & S2 are talking as they start their day. S1 & S2 are sharing thoughts over breakfast.	✓
	Bedtime talk	The speakers are engaging in routine talk right before going to bed.	S1 & S2 are sharing thoughts before going to sleep. S1 & S2 are discussing their day before bed. S1 & S2 are having a chat before bed	✓
	Reminiscing	The speakers are sharing events they experienced together in the past.	S1 & S2 are talking about the good old days. S1 & S2 are reminiscing about past experiences. S1 & S2 are recalling memories they shared.	✓

Table 30: Taxonomy of Speech Events of the category Informal/Superficial Talk.

J Persona Profiles Taxonomy

Underlined is what's relevant to taxonomy and in color-boxes represent what's specific to the language and associated culture and folk psychology in general.

J.1 WELLNESS

Table 31: Taxonomy of WELLNESS category, associated multi-polarised reformulations sentences for the prompt and examples in some languages.

Cat	Sub Category	Sentences	Examples
Symptom	Physical	a physical symptom you have a bodily indication of a health issue you possess a symptom affecting your body	FR: je souffre de rhumatismes articulaires. (I suffer from joint rheumatism.) EN: I'm constantly dealing with <u>back pain</u> .
	Psychiatric	a psychiatric symptom you have a psychological issue you experience an emotional or mental concern you have	FR: j'ai des crises d'angoisse nocturnes. (I have <u>nighttime anxiety attacks</u> .) EN: I have <u>anxiety</u> during big crowds.

continued on next page

Category	Sub Category	Sentences	Examples
Disease	/	a disease you had or currently have a skin condition you had or currently have a digestive respiratory disease you had or currently have	FR: j'ai eu la bronchite à l'âge de 10 ans. (I had <u>bronchitis</u> at age 10.) EN: I'm <u>living</u> with type <u>1</u> diabetes.
↓↓ NEW SUBCATEGORIES FROM TAXONOMY AUGMENTATION ↓↓			
Preventive Care	Check-ups	your health check-ups if any your routine medical exams your preventive health care practices	FR: je fais des échographies cardiaques tous les ans. (I get <u>cardiac ultrasounds</u> every year.) EN: I see my <u>dentist</u> twice a year for cleanings.
	Vaccinations	a vaccine you took or not your vaccination status your stance on vaccines	FR: j'ai reçu ma 3e dose de vaccin contre la grippe. (I received my <u>3rd flu vaccine dose</u> .) EN: I'm <u>not vaccinated</u> against the HPV virus.
	Alternative Medication	your use of herbal remedies your experience with naturopathy your belief in alternative healing methods	FR: j'utilise l'acupuncture pour soulager mon dos. (I use <u>acupuncture</u> to relieve my back pain.) EN: i've used <u>essential oils</u> to relieve anxiety.
Mental Health	Therapy / Counseling	any mental therapy or counseling experiences you had or wish to your mental health treatment your use of psychological services	FR: je vais au psychologue chaque semaine. (I <u>see a psychologist</u> weekly.) EN: I've <u>been in therapy</u> for post-traumatic stress.
Fitness	Exercise Routine	your exercise routine if any how often you work out your physical activity level	FR: j'adore faire du vélo le dimanche matin sur les berges de la seine. (I love riding my bike on Sunday mornings along the <u>banks of the Seine</u> .) EN: I go to the <u>gym</u> three times a week.
Diet	Dietary Restrictions	your food allergies or intolerances if any your adherence to specific dietary plans your special dietary needs	FR: je suis intolérant aux gluten. (I am <u>gluten intolerant</u> .) EN: I have a <u>severe allergy</u> to shellfish.
	Nutritional Habits	your nutritional habits your meal patterns your approach to nutrition	FR: j'adore manger des crêpes bretonnes. (I love eating <u>Breton crepes</u> .) EN: I eat a <u>vegan</u> diet for environmental reasons.
Sleep	Sleep Quality	the amount of sleep you get your experiences with insomnia your methods for improving sleep	FR: je m'endors à 22h00 avec une lecture. (I fall asleep at 10:00 PM with a book.) EN: I've <u>struggled with insomnia</u> since college.
Substance Use	Smoking	your smoking cessation your smoking habits or routine whether you've stopped or kept smoking	FR: je ne fume que des Gauloises. (I only smoke <u>Gauloises</u> .) EN: I used to be a heavy smoker, but quit last year.
	Alcohol Consumption	your alcohol consumption or not your typical drinking patterns your tendency to abstain from alcohol	FR: je bois rarement de bière, je préfère le vin. (I rarely drink beer, I <u>prefer wine</u> .) EN: I don't drink, I'm more of a tea person.
	Drug Use	your experiences with dope use or not your perspective on illegal substance abuse your attitude towards recreational substances	FR: je ne consomme pas de cannabis. (I don't use <u>cannabis</u> .) EN: I think marijuana is overrated.
Medical History	/	significant health events in your past your surgeries past or scheduled if any your family health issues	FR: j'ai eu une opération pour enlever un kyste sur mon pied. (I had a <u>cyst removed</u> from my foot.) EN: I had a tonsillectomy when I was 10.

J.2 PSYCHOGRAPHICS

Table 32: Taxonomy of PSYCHOGRAPHICS category, associated multi-polarised reformulations sentences for the prompt and examples in some languages

Category	Sub Category	Sentences	Examples
Preferences	Movie genre	your favorite type of movie movie genres you avoid the kind of films you enjoy or not	FR: j'ai un faible pour les thrillers. (I have a soft spot for <u>thrillers</u> .) EN: I love watching <u>Marvel superhero movies</u> .
	Movie title	the title of a movie you avoid the movies you love or hate your favorite film	EN: I dislike movies like the <u>Expendables</u> franchise. FR: mon film préféré est amélie poulain. (my favorite movie is <u>Amélie Poulain</u> .)
	Movie director	a movie director whose work you admire your favorite filmmakers a filmmaker you tend to avoid	EN: I dislike Stanley Kubrick's movies, too long. FR: je ne regarde pas les films de luc besson. (I don't watch <u>movies by Luc Besson</u> .)
	Book author	your favorite writer an author you don't enjoy a novelist you admire	FR: Je suis un grand fan de Michel Hoellebecq (I am a big fan of <u>Michel Hoellebecq</u> .) EN: my go-to author is <u>george orwell</u> .
	Book genre	book genres you avoid your preferred or least liked book genre your favorite types of books	EN: i enjoy reading sci-fi novels. FR: je préfère les romans policiers de fredéric dard. (I prefer <u>detective novels</u> by <u>Frederic Dard</u> .)
	Book title	the title of a book you enjoy the most your favorite books a book you dislike	EN: My favorite book is " <u>The Hitchhiker's Guide to the Galaxy</u> ". FR: je préfère "l'étranger" d'camus (I prefer <u>"L'Étranger"</u> by <u>Camus</u> .)
	Show	your favorite television programs the series you enjoy the most a TV show you dislike	FR: je déteste "Koh-Lanta". (I hate <u>"Koh-Lanta"</u> .) EN: i dislike reality TV shows like <u>survivor</u> .

Cat	Sub Category	Sentences	Examples
Preferences	Activity	your leisure activities your favorite outdoor activity your preferred or least liked social activity	FR: j'aime faire du vélo dans les Pyrénées. (I like cycling in the Pyrenees.) EN: i enjoy playing soccer with my friends.
	Season	the time of year you enjoy the most your preferred or least liked season the season you prefer the least	EN: autumn is the worst season for allergies. FR: je déteste l'hiver. (I hate winter.)
	Music instrument	the musical tools you avoid your favorite instruments to play or listen to your preferred or least liked music instrument	EN: i enjoy listening to acoustic guitar music. FR: je déteste le saxophone (I hate the saxophone.)
	Music genre	your preferred or least liked music genre music genres you dislike the types of music you enjoy	EN: i love indie rock music. FR: je n'aime pas le jazz. (I don't like jazz.)
	Music artist	the musician you admire your most loved singers or bands the artist you avoid listening to	FR: j'dore les albums de Claude François. (I love the albums of Claude François.) EN: i listen to the 1975 on repeat.
	Color	the colors you dislike the color you prefer the least your favorite color	EN: I'm not a fan of the color green. FR: mon coloris préféré est le vert foncé (my favorite color is dark green.)
	Animal	an animal you find fascinating or repellent your favorite or least liked animals your interest in wildlife	EN: I'm fascinated by the behavior of dolphins in the wild. FR: les crocodiles me font peur. (Crocodiles scare me.)
	Location / Place	locations you dislike locations you avoid places you love to visit	EN: i love exploring the mountains of colorado. FR: Je adore passer mes week-ends à la plage de Saint-Tropez. (I love spending my weekends at Saint-Tropez beach.)
	Sport	your preferred or least liked sport the sports you enjoy a sport you don't like	EN: i'm a huge fan of soccer. FR: J'adore jouer au pétanque avec mes amis tous les dimanches. (I enjoy playing pétanque with my friends every Sunday.)
	Food	the cuisines you enjoy or not foods you avoid your favorite dishes	EN: i love eating fish and chips from the seaside. FR: je mange souvent des croissants. (I often eat croissants.)
	Drink	your favorite drink your preferred or least liked drink beverages you avoid	EN: i hate the taste of Earl Grey tea. FR: je déteste le café glacé. (I hate iced coffee.)
	Media genre	the kind of media do you find most engaging your preferred or least liked media genre the media genres you avoid	EN: i enjoy true crime podcasts. FR: je déteste les émissions de télé-réalité (I hate reality TV shows.)
	Education Methods	your learning style your preferred or least liked teaching methods how you best learn	FR: j'adore l'apprentissage en classe mixte. (I love learning in "classe mixte".) EN: I learn best through hands-on experience.
	Favorite Apps	your favorite apps if any your go-to digital tools and platforms the most useful app of yours	FR: mon application préférée est Waze. (my favorite app is Waze.) EN: i'm active on instagram and tiktok mostly.
Personal Characteristics	Physical Attribute	a specific aspect of your physical appearance a particular aspect of your body you don't like one of your physical traits, noticeable or subtle	EN: i have a scar above my left eyebrow. FR: J'ai un tatouage de la tour Eiffel sur mon épaule gauche. (I have a tattoo of the Eiffel Tower on my left shoulder.)
	/	any dimension of your personality your level of empathy/agreeableness or Neuroticism your self-awareness or conscientiousness	EN: i'm a bit of a control freak. FR: je suis un peu trop sensible. (I'm a bit too sensitive.)
	Decision -Making Style	your decision-making style: intuitive, rational or collaborative how you make decisions your method of making choices	EN: I prefer to think things through before acting impulsively. FR: je prends des décisions en écoutant mes émotions. I make decisions by listening to my emotions.
	Communication Style	your preferred communication methods how you express yourself your verbal or non-verbal communication style	EN: i'm a naturally quiet person in large groups. FR: je préfère les conversations en direct. (I prefer face-to-face conversations.)
	Problem Solving	your analytical thinking skills your creativity in finding solutions or not your peacekeeping abilities or not	EN: i'm not very good at mediating conflicts. FR: je suis très calme en situation de crise. (I'm very calm in crisis situations.)
	Resilience	your adaptability skills your ability to overcome adversity or not your coping strategies	EN: i struggle to set realistic goals for myself sometimes. FR: je suis capable de changer de projet rapidement. (I am capable of quickly switching projects.)
	Creativity	your imagination and originality your artistic skills your innovative thinking	EN: i'm decent at painting watercolors FR: je compose des chansons sur mon ukulélé. (I compose songs on my ukulele.)
	Core Values	your core values or moral standpoints your ethical principles your fundamental beliefs	EN: My moral compass is centered around being honest FR: Je tiens à mon indépendance et à mes principes. (I value my independence and my principles.)
	Cognitive Abilities	your intellectual capabilities your level of attention your memory abilities	EN: I have trouble focusing during long meetings. FR: j'ai une mémoire incroyable pour les chansons. (I have an incredible memory for songs.)

continued on next page

Cat	Sub Category	Sentences	Examples
Personal Characteristics	Financial Awareness	your saving habits how you manage your finances your investing knowledge	EN: i'm learning about real estate investing, hoping to flip a house. FR: je fais l'épargne pour acheter un appartement à Nice. (I am saving to buy an apartment in Nice .)
	Spending Habits	your shopping tendencies your consumer behavior your spending habits	EN: i always look for sales when shopping at Target . FR: je préfère payer cash pour éviter les intérêts. (I prefer paying cash to avoid interest.)
	Lifestyle	your typical day your day-to-day life your regular schedule	EN: my typical day starts with a 5am jog along the beach FR: je prends le métro tous les jours. (I take the subway every day.)
	Social Connections	your social bonding skills your network of friends your clubs or associations if any	FR: je suis très lié avec mes copains de la fac à strasbourg. I am very close with my friends from university in Strasbourg . EN: i'm terrible at remembering birthdays.
↓↓ NEW CATEGORY AND ENTITIES FROM TAXONOMY AUGMENTATION ↓↓			
Interests	Technology	your enthusiasm for technology your adaptability to new technology or not your awareness of cyber threats	FR: je suis très actif sur mon compte Instagram. (I am very active on my Instagram account .) EN: i'm not tech-savvy, i still use a flip phone .
	Hobby and passions	your interests in art if any any of your hobbies or passions your various interests	EN: i enjoy hiking in the pacific northwest . FR: je suis passionné par la peinture de Claude Monet. (I am passionate about the paintings of Claude Monet .)
	/	your environmental advocacy efforts if any your commitment to sustainability your participation in environmental campaigns	EN: i'm passionate about reducing plastic waste in our oceans. FR: j'appuie les initiatives pour protéger la biodiversité. (I support initiatives to protect biodiversity.)
	Environment	reusable products you incorporate into your daily life your zero waste efforts your eco-friendly choices and practices	EN: i always wear second-hand clothes , it's more sustainable. FR: j'ai un composteur dans mon jardin. (I have a composter in my garden.)
Goals	Carbon Footprint	your use of renewable energy your resource consumption your minimalist lifestyle to save the planet	EN: i generate a lot of paper waste at home FR: je préfère utiliser le bus. (I prefer using the bus.)
	Travel	your interest in travelling memorable trips you've taken countries you want to explore	FR: j'ai visité le Mont-Saint-Michel à l'âge de 12 ans. (I visited the Mont-Saint-Michel at the age of 12.) EN: i've traveled to europe many times.
	Personal Goals	your personal or life goals one of your key life purposes any extrinsic or intrinsic motivation of yours	FR: je suis motivé pour réussir mon bac. (I am motivated to succeed my baccalaureate .) EN: I aspire to start my own business.
Influences	Role Models	someone that inspires you someone whose life you admire a hero in your eyes	EN: my hero is steve jobs . FR: Je admire les réalisations de l'astronaute Thomas Pesquet. (I admire the achievements of astronaut Thomas Pesquet .)

J.3 DEMOGRAPHICS

Table 33: Taxonomy of DEMOGRAPHICS category, associated sentences for the prompt and examples in randomly selected languages

Cat	Sub Category	Sentences	Examples
Location	Residence	your city or country of residence your present hometown where you currently live	FR: je habite actuellement à Lyon, dans le quartier de roix-Rousse. (I currently live in Lyon , in the Croix-Rousse neighborhood.) EN: I call Melbourne, Australia home.
	Birthplace	where you were born your city or country of birth your childhood hometown	FR: née à Marseille, j'ai une âme méditerranéenne. (Born in Marseille , I have a Mediterranean soul.) EN: i grew up in austin, texas .
	Nationality	the country you are a citizen of your nationality the nation you belong to	FR: je suis belge. (I am Belgian .) EN: I'm a true British person at heart.
Employment	Company Name	official name of the organization you work for the company you are employed with the business you work for	FR: je suis employé à la SNCF. (I work for SNCF .) EN: I'm employed by the British Museum .
	Workplace	your work environment your office setting the place where you work	FR: je travaille au siège social de la société BPCE. (I work at the headquarters of BPCE .) EN: i work in a coffee shop in the financial district .
	Profession	your current or previous profession your job skills or certifications the field you work in	FR: je suis enseignant de français en chine. (I work as a French teacher in China.) EN: i'm a part-time yoga instructor.
	Job Status	whether you are currently employed or not if you are working or seeking employment your job situation	FR: je suis à la retraite. (I am retired.) EN: i'm currently between jobs after a layoff.

continued on next page

Cat	Sub Category	Sentences	Examples
Employment	Career Path	your professional milestones the trajectory of your career the steps you've taken in your career	FR: j'ai créé une boutique de mode à Montmartre après avoir étudié à la ESMOD. (I created a fashion boutique in Montmartre after studying at ESMOD .) EN: I've worked as a software engineer for five years in Silicon Valley .
	Job Fulfillment	your career happiness how satisfied you are with your job your professional bliss	FR: j'adore mon travail d'enseignant. (I love my job as a teacher.) EN: I'm not satisfied with my current job, I want more challenge.
	Motivations and Goals	your professional goals what drives you at work your work aspirations	FR: j'aimerais écrire un livre sur la Révolution française. (I would like to write a book about the French Revolution .) EN: my ultimate goal is to work for the national park service .
	Work-Life Balance	your work flexibility how you manage work-life balance your methods for balancing work and personal life	FR: je travaille en freelance pour avoir plus de liberté. (I work freelance to have more freedom.) EN: i prioritize work during the week, family on weekends.
	Remote Work	your home office setup if any your telecommuting experience if any your remote work best or worst practices	FR: j'ai travaillé à domicile pendant un an. (I worked from home for a year.) EN: i get distracted by notifications when remote working.
	Network	your professional network your mentee if any your industry connections	FR: je suis mentoré par mon collègue, Pierre Dupont. (I am mentored by my colleague, Pierre Dupont .) EN: i'm trying to expand my network on linkedin.
Education	Degree	the current degree you pursue the degree you have earned the level of education you are working towards	FR: j'ai obtenu un diplôme en sociologie à l'Université de Paris. (I have earned a degree in sociology from the University of Paris .) EN: i have a master's in finance from nyu .
	Degree subject	your field of study your academic discipline your degree major	FR: j'ai obtenu un master en marketing à l'ESSEC. (I got a master in marketing at the ESSEC .) EN: My degree is in linguistics from the University of Michigan .
	School Name	your alma mater the school you graduated from the school institution you attend	EN: i'm enrolled in a master's program at UCLA . FR: j'ai étudié à l'École normale supérieure de Lyon. (I studied at the École Normale Supérieure de Lyon .)
	School Status	your school status: student, alumni, etc if you have graduated whether you are currently a student	EN: i graduated summa cum laude . FR: je suis diplômé en sciences politiques de Sciences Po Paris. (I'm graduated in political science from Sciences Po Paris .)
	School type	the type of school you attend whether you attend a public or private school the category of your school	FR: j'ai étudié à l'université Paris-Sorbonne. (I studied at the Paris-Sorbonne university.) EN: i'm a student at a community college .
	↓↓ NEW ENTITIES FROM TAXONOMY AUGMENTATION ↓↓		
Family Status	Achievements	any honor or award you wished to receive at school your academic accomplishments the awards you earned in school	FR: j'aimerais recevoir le prix littéraire des lycéens. (I would love to receive the literary prize for high school students.) EN: I wish I had won a Pulitzer Prize for my writing.
	Extracurricular Activities	any club or association you were involved in at school projects you worked on outside of class your participation in school sports	FR: j'ai été vice-président du club de rugby. (I was vice-president of the rugby club .) EN: i played basketball in college , but got injured.
	Spoken Languages	other languages or dialects that you speak or learn to the languages you are fluent in additional languages you speak	FR: je parle français et Breton. (I speak French and Breton .) EN: I'm learning Spanish to communicate with my clients .
	Workshops / Seminars	workshops or seminars you attended or wish to training programs you completed seminars where you presented some work	FR: j'ai présenté mon projet à la station f. (I presented my project at Station F .) EN: i presented my startup idea at the techcrunch conference.
	Siblings	your brothers or sisters the number of siblings you have your family members	FR: j'ai deux sœurs qui vivent en province. I have two sisters who live in the provinces. EN: I have a twin sister named Brittany .
	Children	your children if any the children in your family if any your offspring	FR: j'ai deux jeunes filles, Sophie et Emma. (I have two young daughters , Sophie and Emma .) EN: I'm a single parent with a 10-year-old son who loves Legos
Possession	Animal	your pet if you have one an animal companion you want an animal you possess or wish to	EN: my dream pet is a capybara . FR: j'ai toujours voulu un chien berger allemand. (I always wanted a german shepherd dog .)
	Vehicle	your dream car or vehicle a vehicle you own or plan to buy a means of transportation you possess	EN: i own a 1969 ford mustang . FR: ma voiture actuelle est une renault clio. (My current car is a Renault Clio .)
	↓↓ NEW ENTITY FROM TAXONOMY AUGMENTATION ↓↓		
	Tech Device	your favorite gadget your go-to tech tool an electronic device you own or desire	EN: i'm really interested in getting a 3d printer . FR: J'ai un ordinateur portable Apple MacBook Pro que j'adore. (I have a MacBook Pro laptop that I love.)

continued from previous page

Category	Sub Category	Sentences	Examples
Marital Status	/	your marital status if you have a spouse or partner whether you are married, single or divorced	EN: i'm divorced. FR: je suis marié avec une Goldenrod femme originaire de Lyon. (I'm married to a woman originating from Lyon.)
Name	/	the name you go by the name you are known as your full name	EN: my name is Maxwell Thompson. FR: je m'appelle Sophie Lefebvre. (My name is Sophie Lefebvre.)
Age	/	how old you are your birth year the number of years you have lived	EN: i was born in 1992. FR: je suis quarantenaire. (I am quarantenaire.)
Gender	/	the gender you identify as your preferred pronouns your gender	EN: i prefer they/them pronouns. FR: je suis une femme. (I am a woman.)
↓↓ NEW CATEGORIES AND ENTITIES FROM TAXONOMY AUGMENTATION ↓↓			
Ethnicity	/	your cultural background, heritage, or ancestry your pride in your culture your ethnic background	EN: I identify as a proud Brit with Scottish heritage. FR: je suis originaire de la Martinique. (I'm native of Martinique.)
Religion / Spirituality	/	your religion or beliefs your faith or lack thereof your spiritual views	EN: i'm a devout Christian, but i don't attend church often. FR: Je suis agnostique et ne pratique pas de religion. (I am agnostic and do not practice religion.)
Socioeconomic Status	Housing Status	your housing status or living arrangements the type of housing you live in your home environment	EN: my apartment has a view of the brooklyn bridge. FR: j' habite dans un village de la Loire-Atlantique. (I live in a village in Loire-Atlantique.)
	Income Level	your social standing your economic position your income level, wealth, or social class	EN: i've been in debt since college. FR: je gagne juste suffisamment pour voyager. (I earn just enough to travel.)

K Per Language Detailed Automatic Analysis

K.1 High-Resource Languages

K.1.1 TURKISH

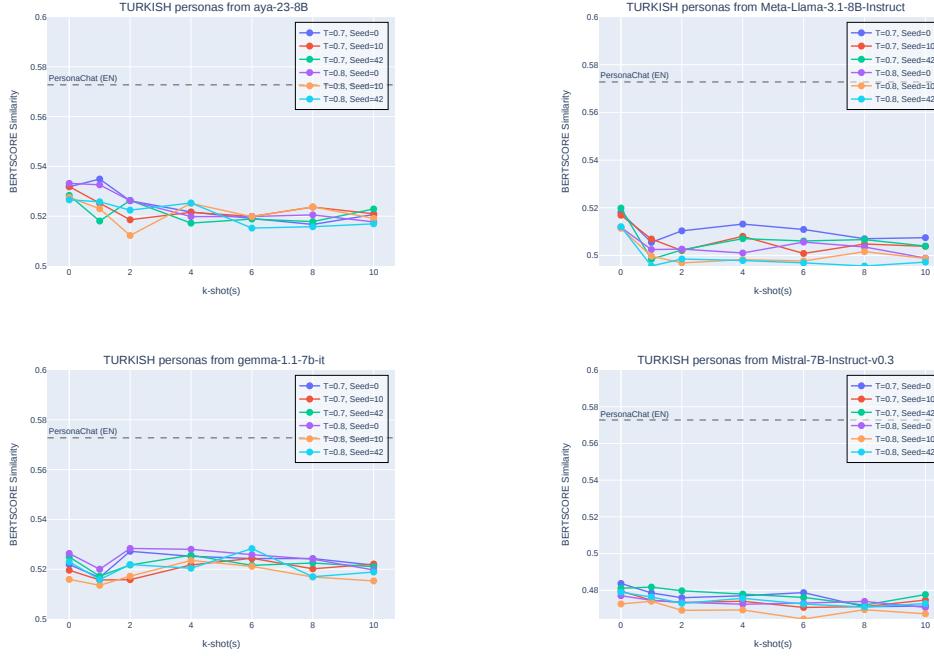


Figure 7: Detailed BERTSCORE for Turkish Personas in different generation configurations for the different models

K.1.2 ENGLISH



Figure 8: Detailed BERTSCORE for English Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.1.3 RUSSIAN



Figure 9: Detailed BERTSCORE for Russian Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.1.4 GERMAN



Figure 10: Detailed BERTSCORE for German Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.1.5 JAPANESE

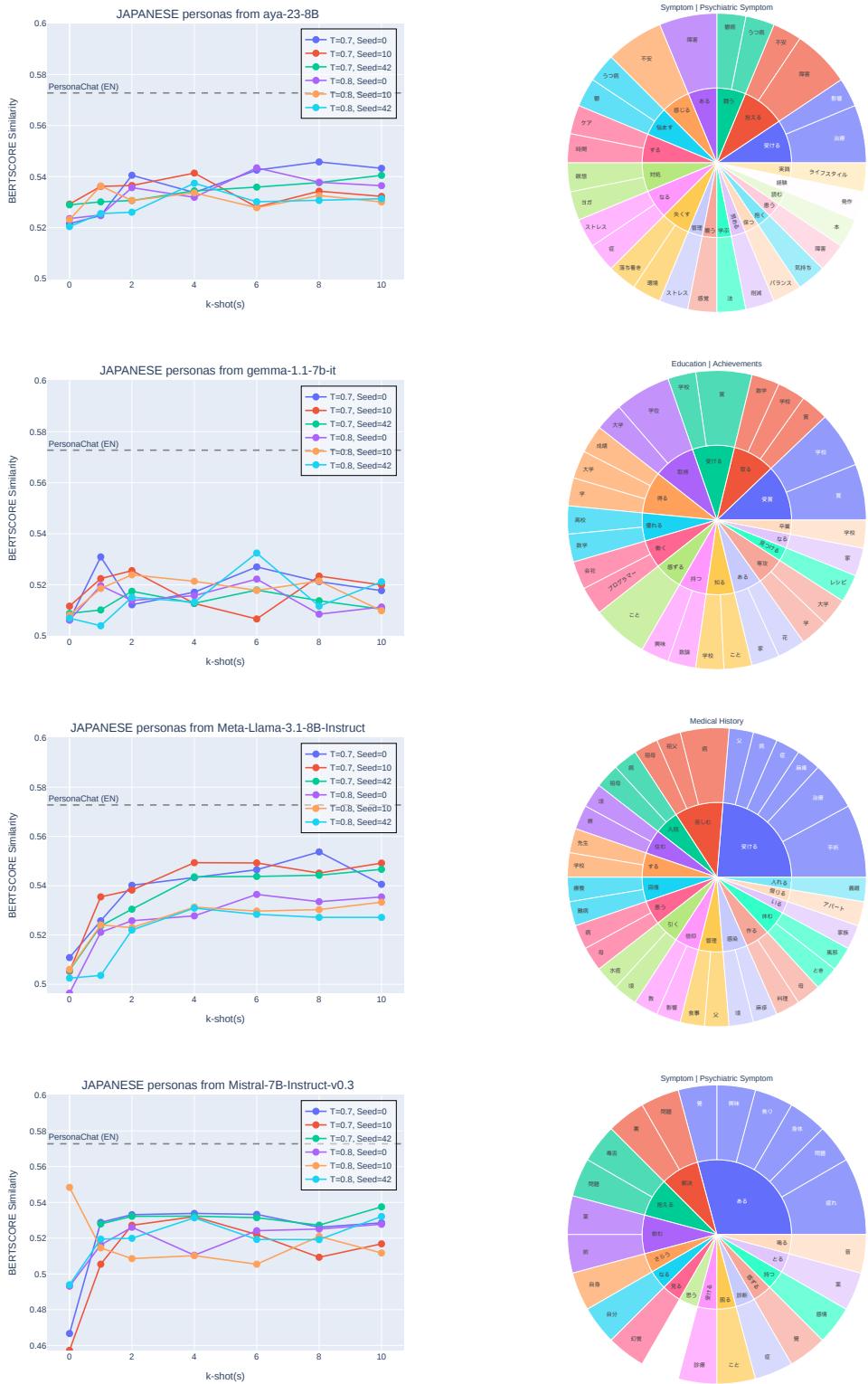


Figure 11: Detailed BERTSCORE for Japanese Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.1.6 SPANISH

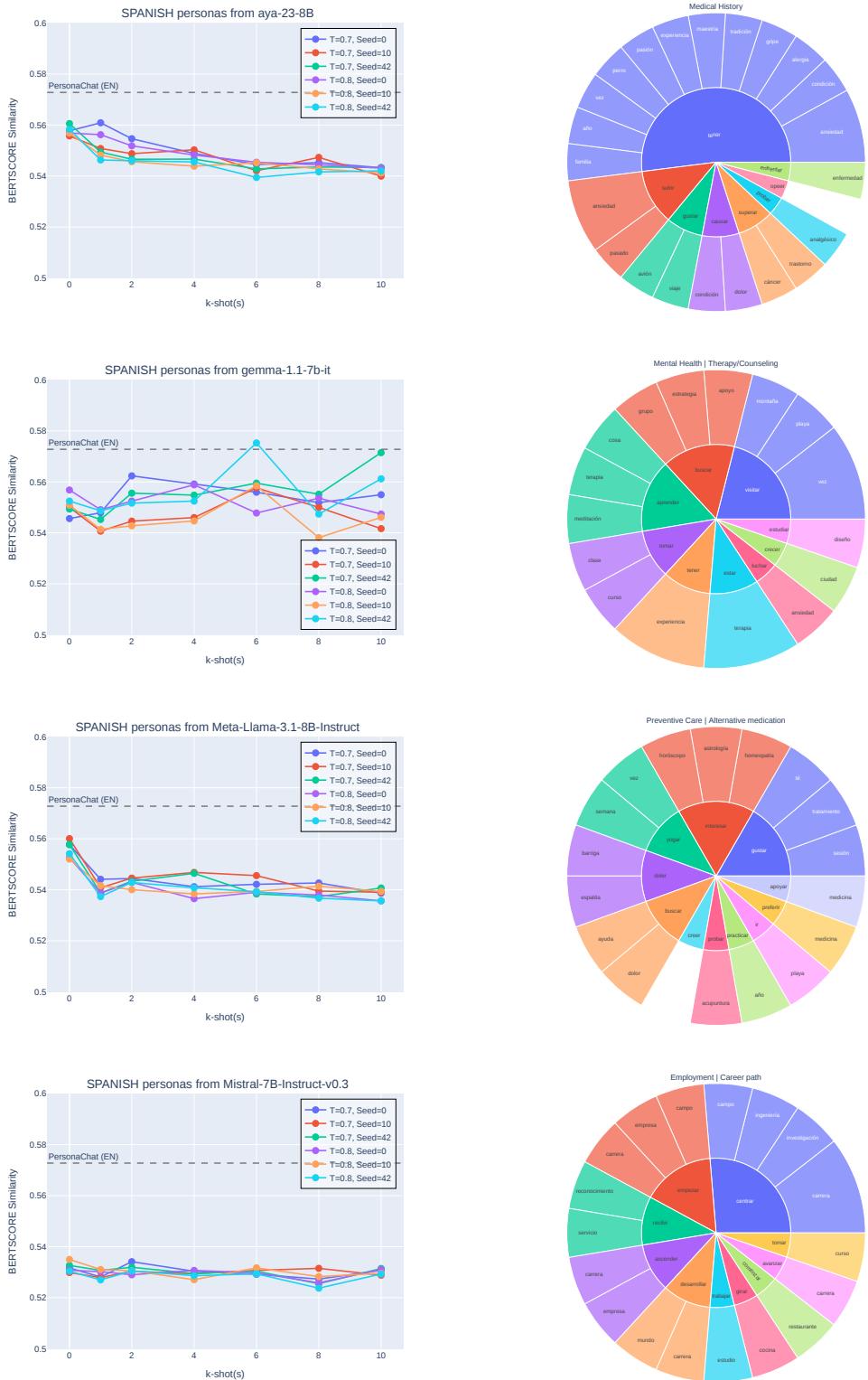


Figure 12: Detailed BERTSCORE for Spanish Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.1.7 CHINESE

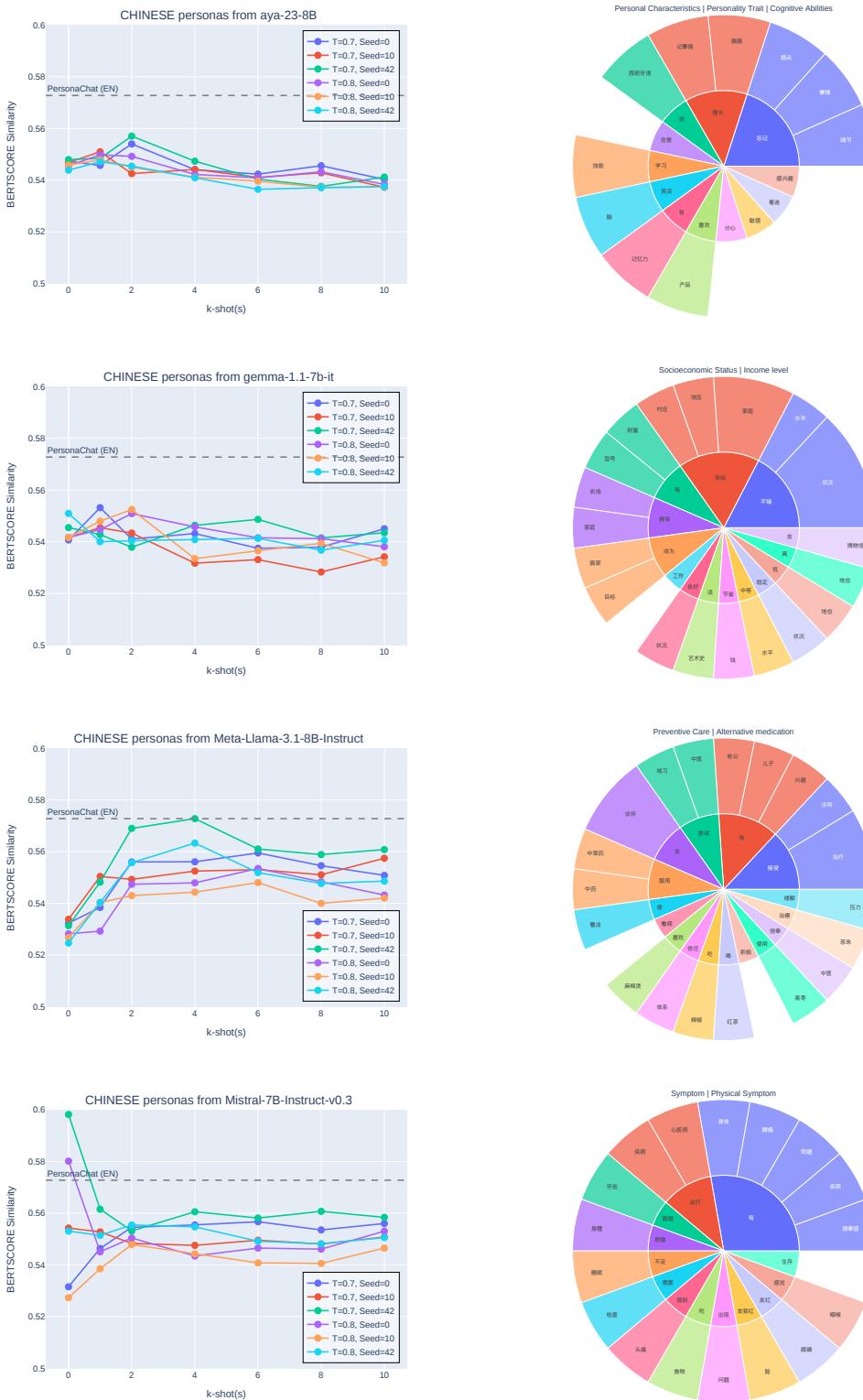


Figure 13: Detailed BERTSCORE for Chinese Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.1.8 FRENCH



Figure 14: Detailed BERTSCORE for French Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.1.9 ITALIAN



Figure 15: Detailed BERTSCORE for Italian Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.1.10 DUTCH



Figure 16: Detailed BERTSCORE for Dutch Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.1.11 PORTUGUESE



Figure 17: Detailed BERTSCORE for Portuguese Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.1.12 POLISH

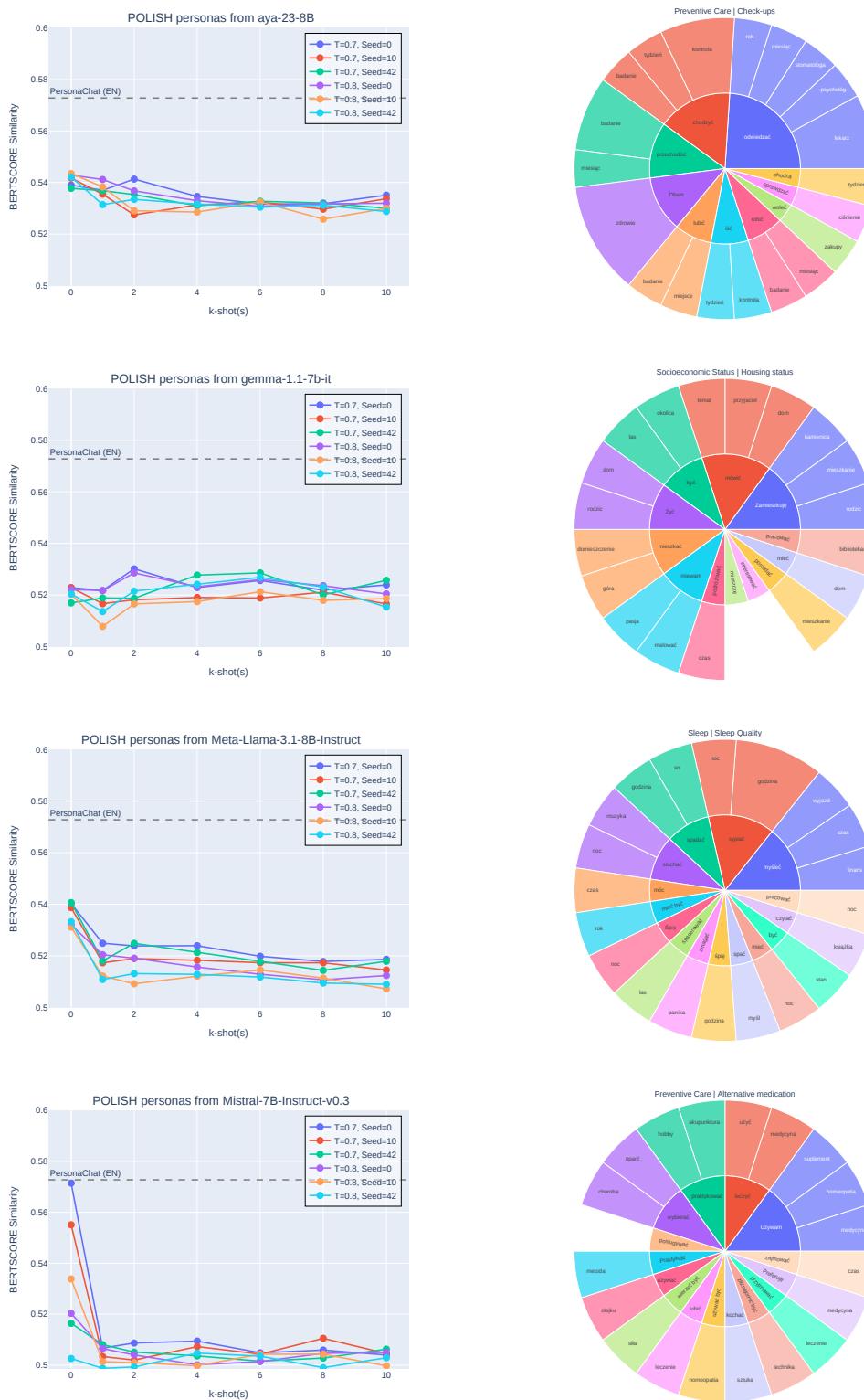


Figure 18: Detailed BERTSCORE for Polish Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.2 Medium-Resource Languages

K.2.1 VIETNAMESE

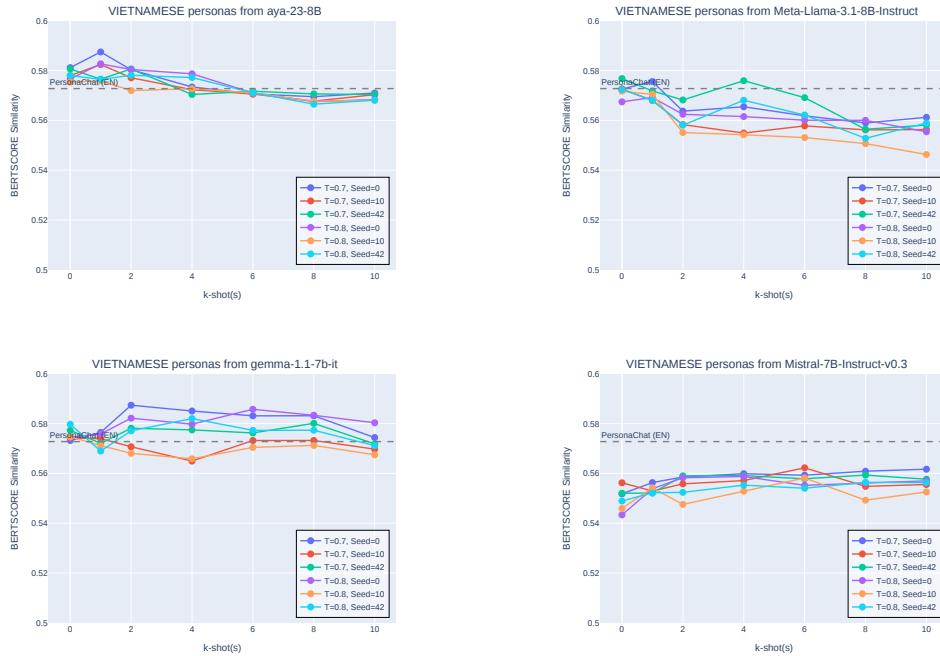


Figure 19: Detailed BERTSCORE for Vietnamese Personas in different generation configurations for the different models

K.2.2 INDONESIAN

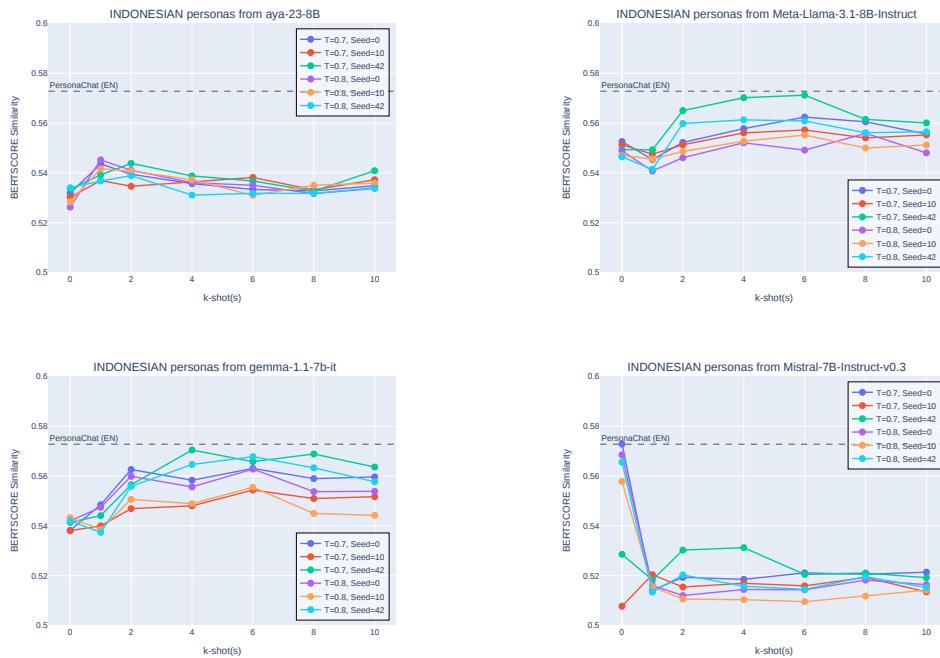


Figure 20: Detailed BERTSCORE for Indonesian Personas in different generation configurations for the different models

K.2.3 KOREAN

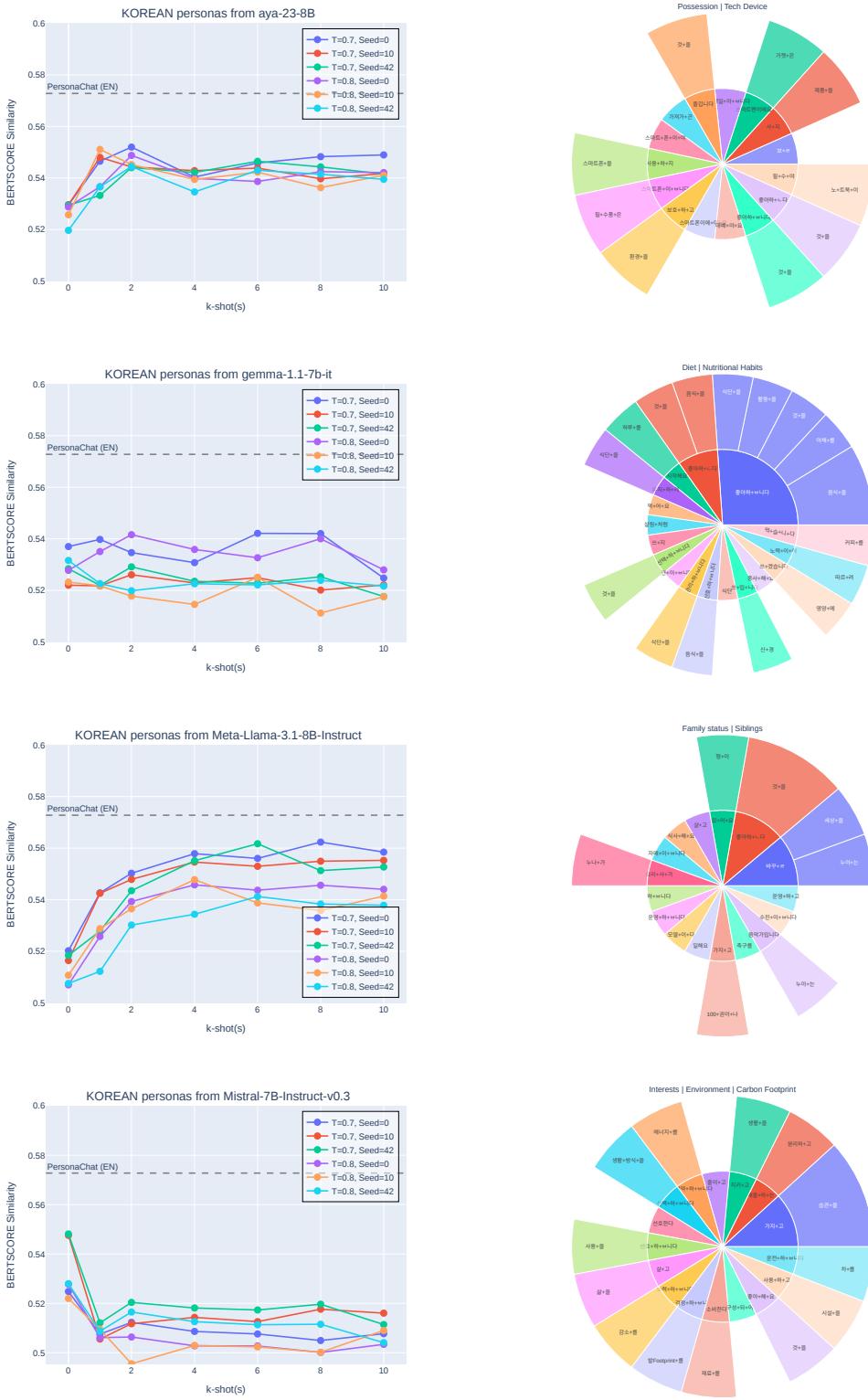


Figure 21: Detailed BERTSCORE for Korean Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.2.4 SWEDISH



Figure 22: Detailed BERTSCORE for Swedish Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.2.5 ARABIC

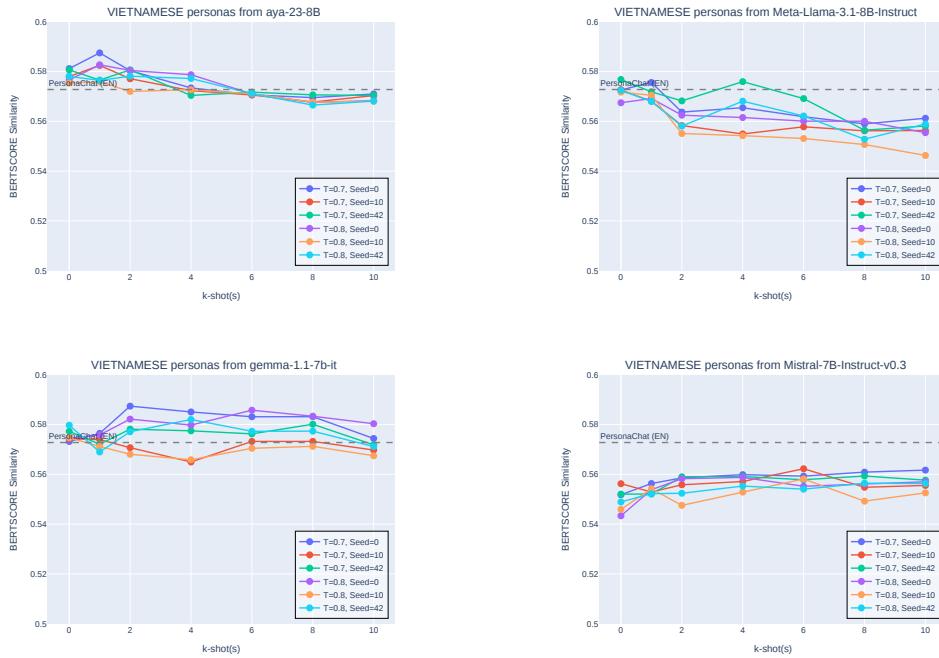


Figure 23: Detailed BERTSCORE for Arabic Personas in different generation configurations for the different models

K.2.6 HUNGARIAN

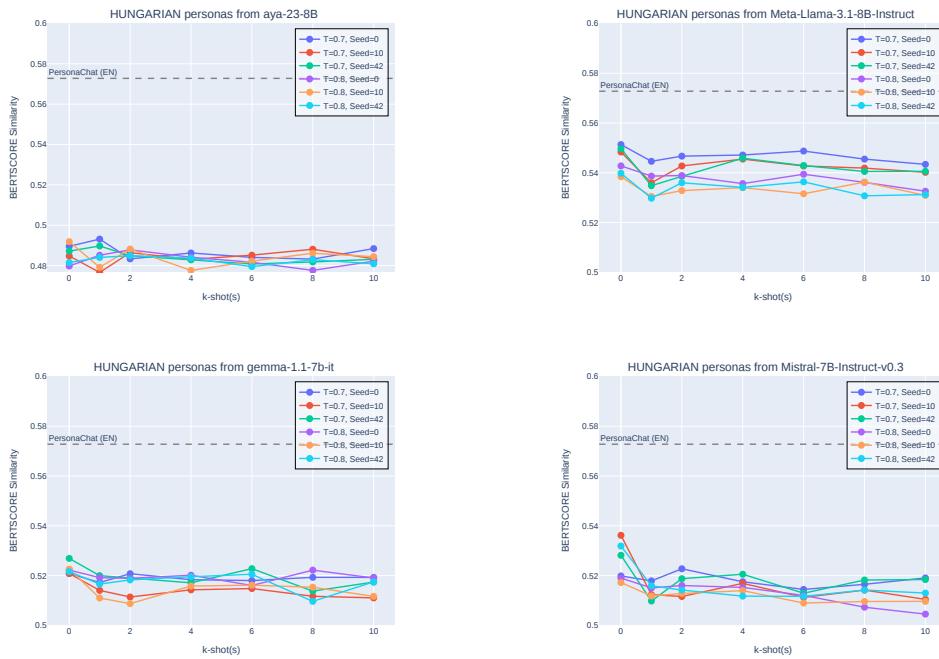


Figure 24: Detailed BERTSCORE for Hungarian Personas in different generation configurations for the different models

K.2.7 GREEK

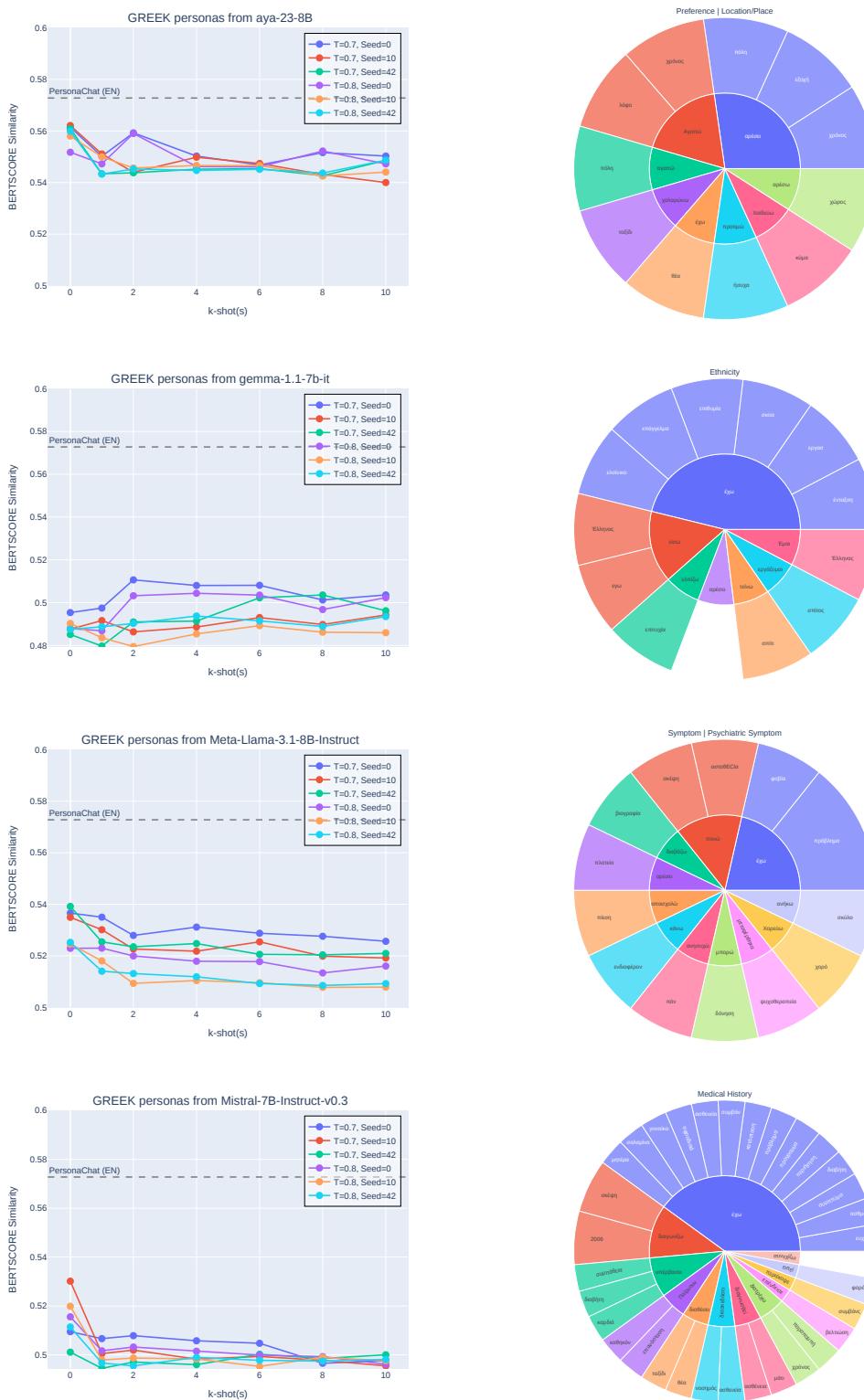


Figure 25: Detailed BERTSCORE for Greek Persons in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.2.8 UKRAINIAN



Figure 26: Detailed BERTSCORE for Ukrainian Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.2.9 DANISH

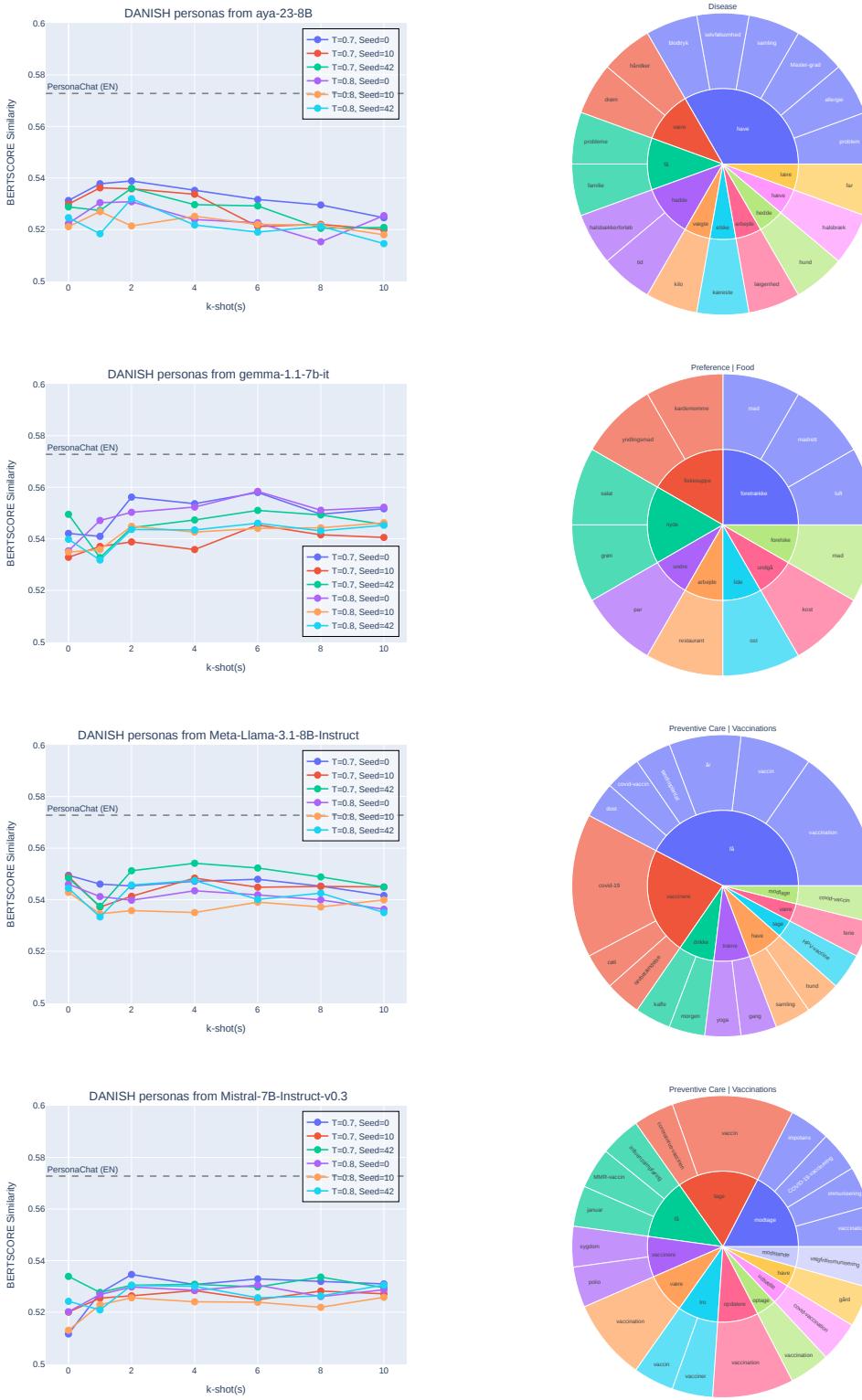


Figure 27: Detailed BERTSCORE for Danish Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.2.10 FINNISH

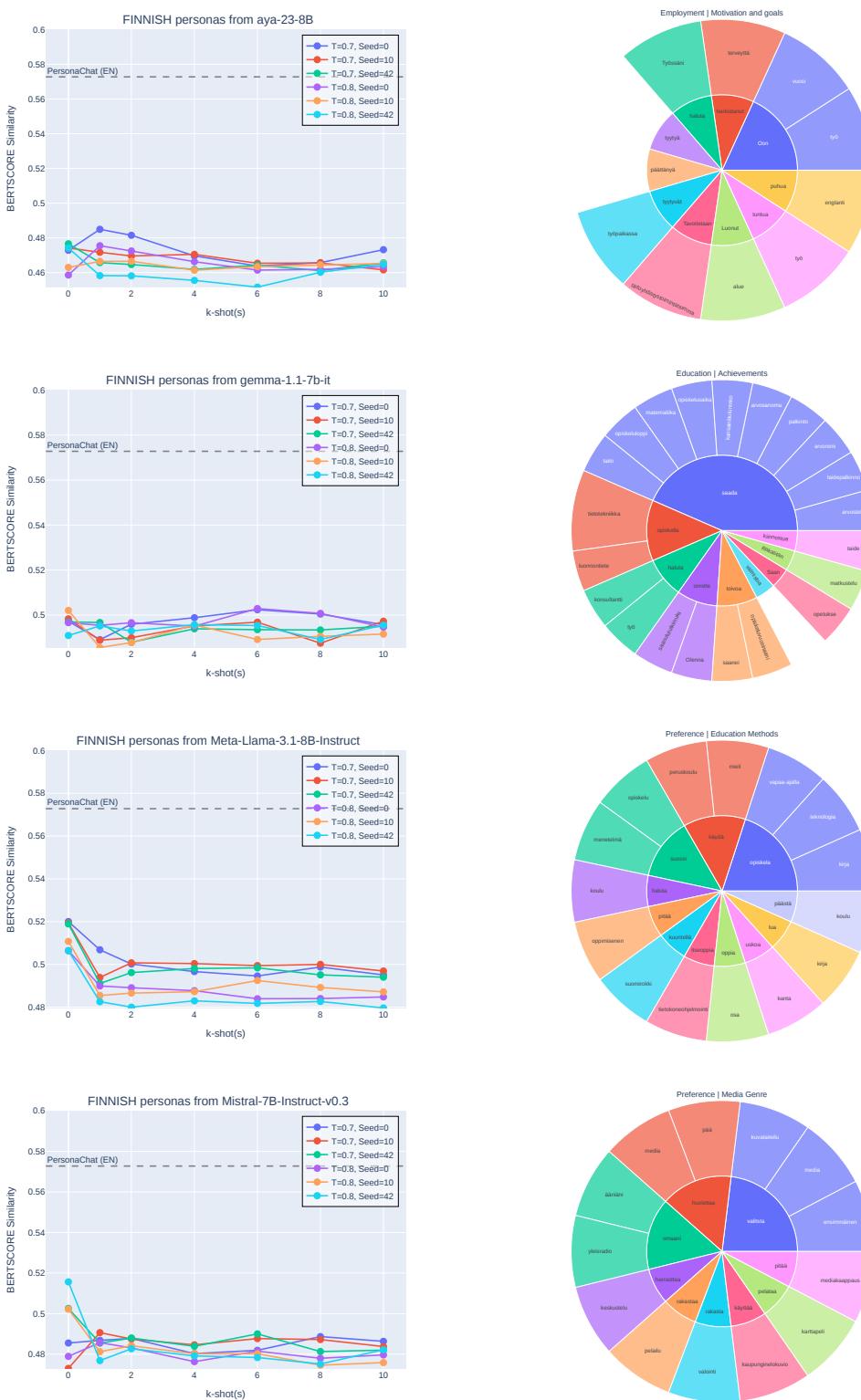


Figure 28: Detailed BERTSCORE for Finnish Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.2.11 CROATIAN

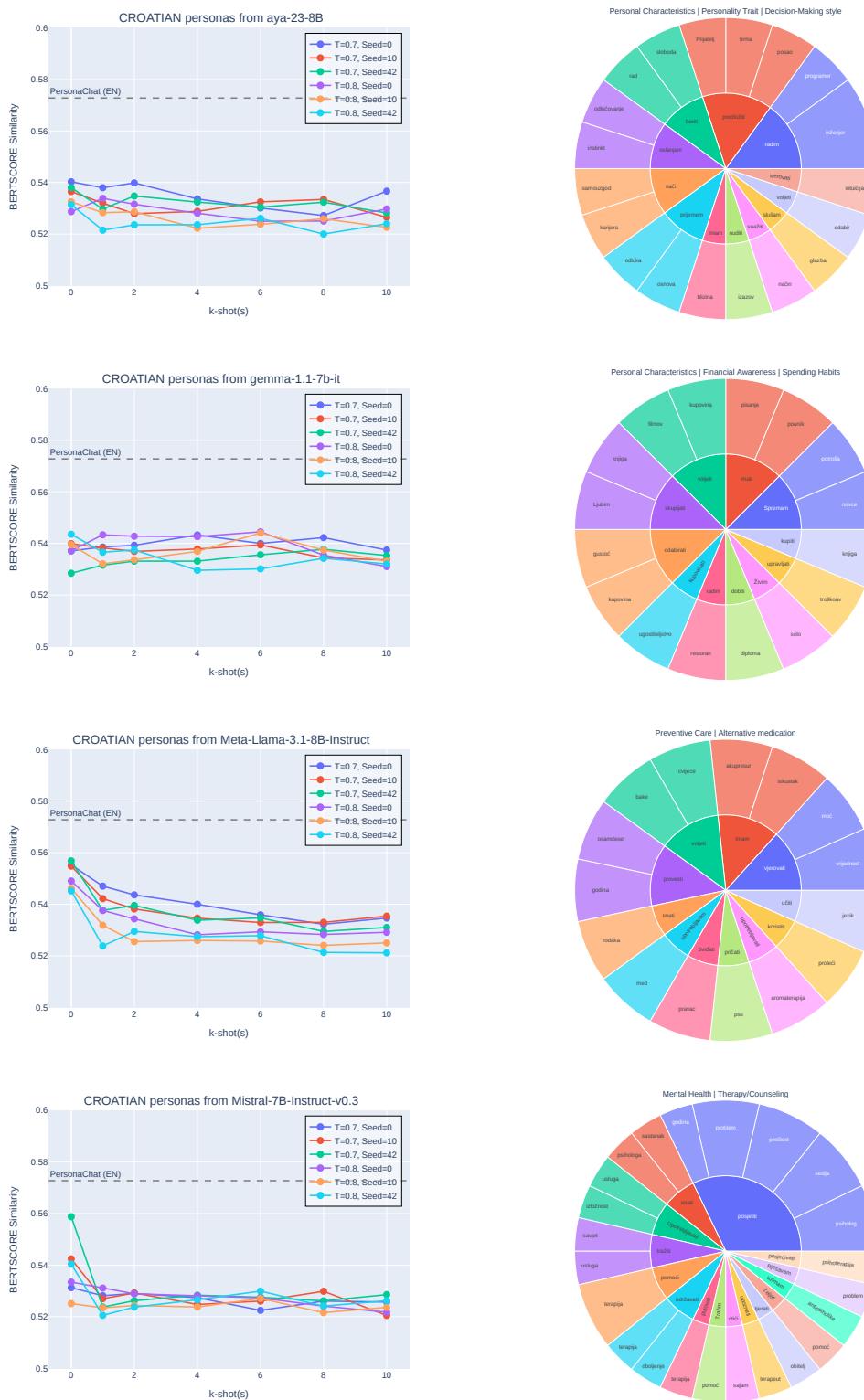


Figure 29: Detailed BERTSCORE for Croatian Personas in different generation configurations and Sunburst charts of personas taxonomy entities with most root verbs and associated object noun for the different models

K.2.12 THAI*

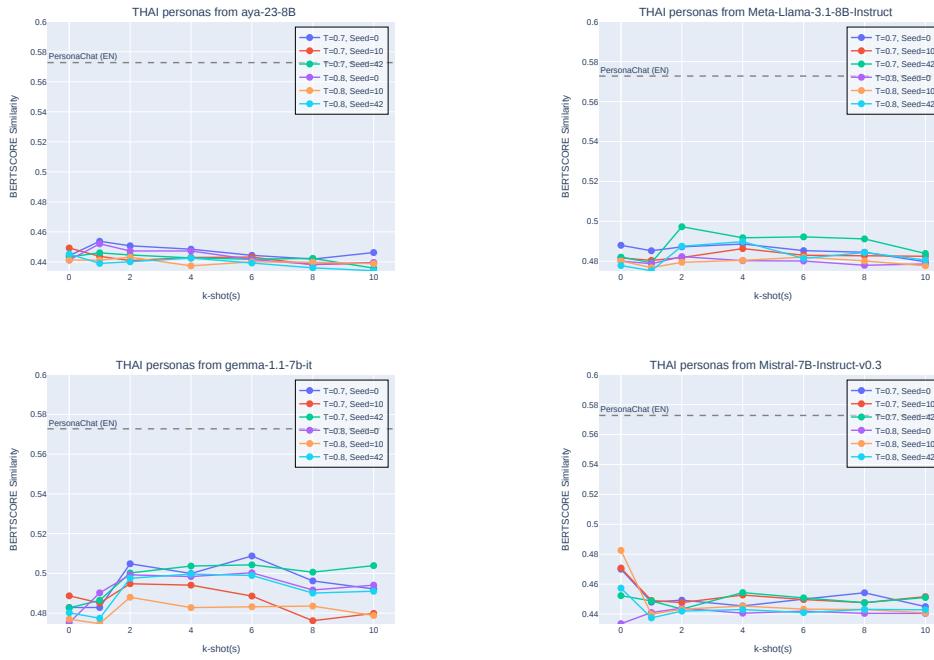


Figure 30: Detailed BERTSCORE for Thai Personas in different generation configurations for the different models

K.2.13 HINDI

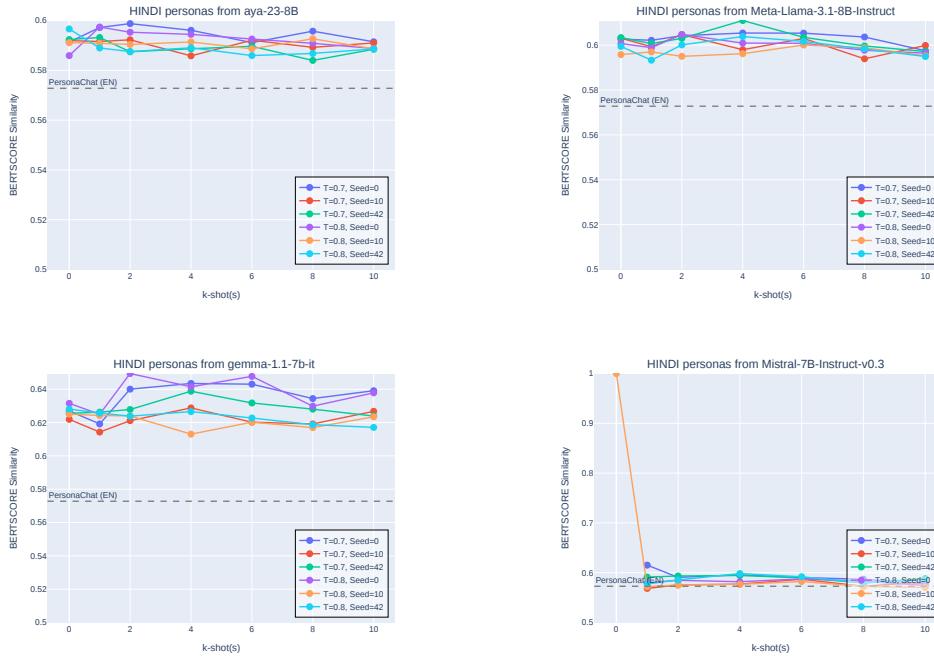


Figure 31: Detailed BERTSCORE for Hindi Personas in different generation configurations for the different models

K.2.14 BENGALI

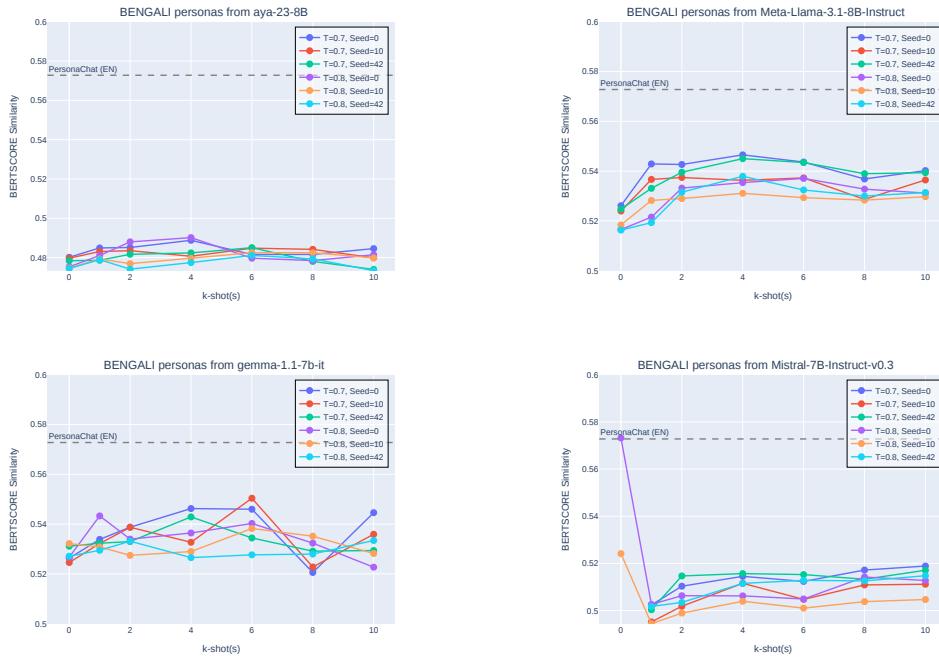


Figure 32: Detailed BERTSCORE for Bengali Personas in different generation configurations for the different models

K.3 Low-Resource Languages

K.3.1 AFRIKAANS

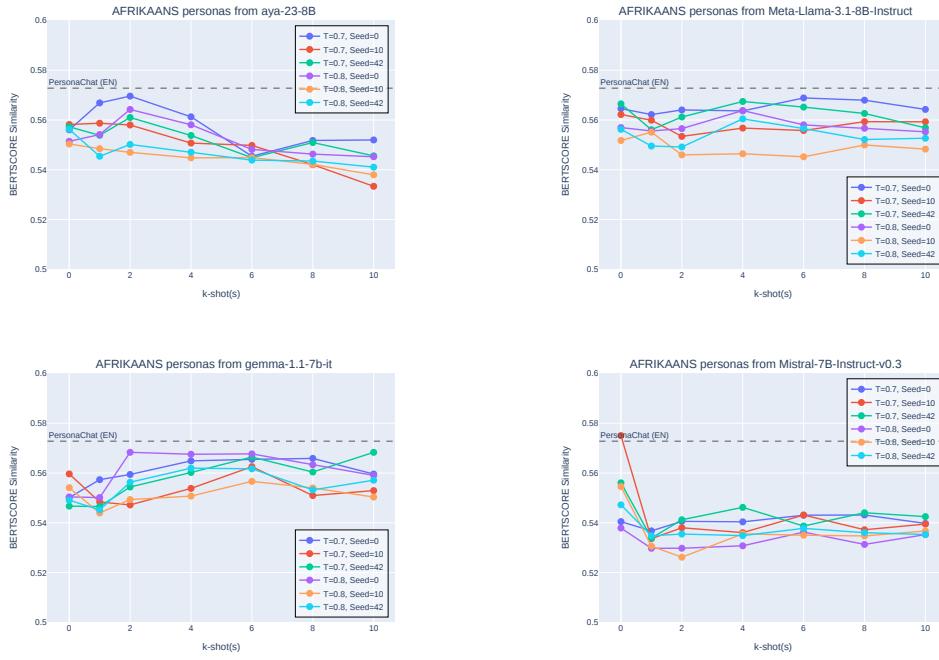


Figure 33: Detailed BERTSCORE for Afrikaans Personas in different generation configurations for the different models

K.3.2 SWAHILI

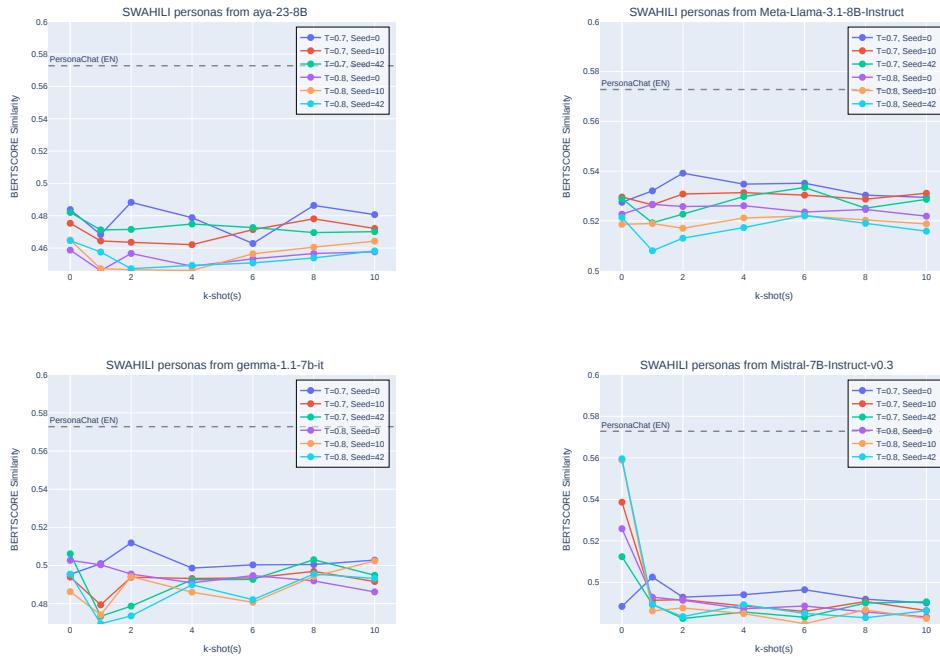


Figure 34: Detailed BERTSCORE for Swahili Personas in different generation configurations for the different models

K.3.3 YORUBA

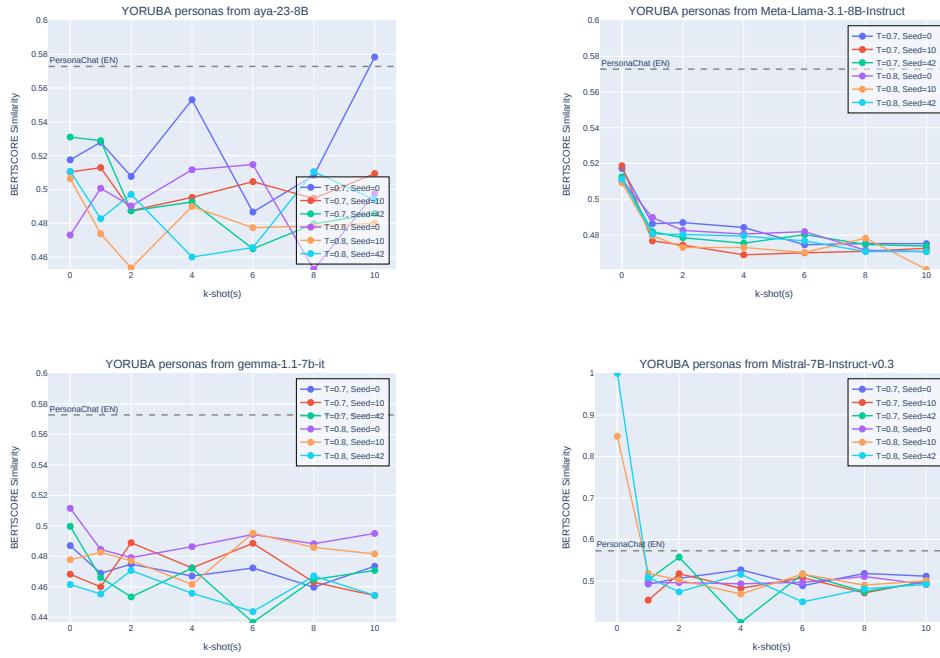


Figure 35: Detailed BERTSCORE for Yoruba Personas in different generation configurations for the different models