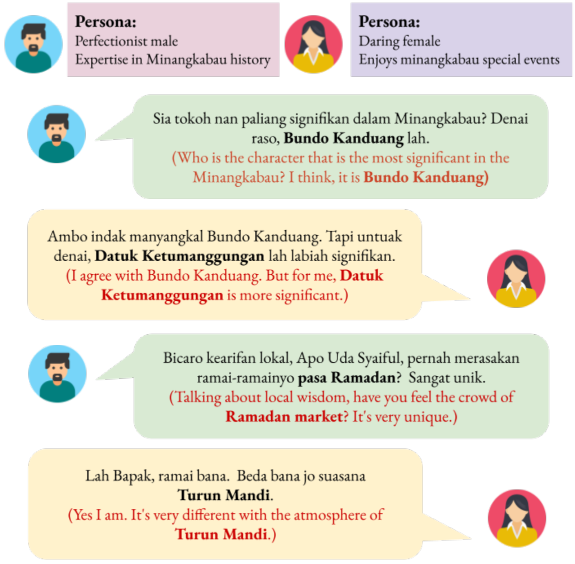
SEADIALOGUES: A Multilingual Culturally Grounded Multi-turn Dialogue Dataset on Southeast Asian Languages

Anonymous ACL submission

Abstract

Although numerous datasets have been developed to support dialogue systems, most existing chit-chat datasets overlook the cultural nuances inherent in natural human conversations. To address this gap, we introduce a culturally grounded dialogue dataset centered on Southeast Asia, a region with over 700 million people and immense cultural diversity. Our dataset features dialogues in eight languages from six Southeast Asian countries, many of which are low-resource despite having sizable speaker populations. To enhance cultural relevance and personalization, each dialogue includes persona attributes and two culturally grounded topics that reflect everyday life in the respective communities. Furthermore, we release a multi-turn dialogue dataset to advance research on culturally aware and human-centric large language models, including conversational dialogue agents.



Persona:
Perfectionist male
Expertise in Minangkabau history

Persona:
Daring female
Enjoys minangkabau special events

Sia tokoh nan paliang signifikan dalam Minangkabau? Denai raso, **Bundo Kanduang** lah.
(Who is the character that is the most significant in the Minangkabau? I think, it is **Bundo Kanduang)**

Ambo indak manyangkal Bundo Kanduang. Tapi untuak denai, **Datuk Ketumanggungan** lah labiah signifikan.
(I agree with Bundo Kanduang. But for me, **Datuk Ketumanggungan** is more significant.)

Bicaro kearifan lokal, Apo Uda Syaiful, pernah merasakan ramai-ramainyo **pasa Ramadan**? Sangat unik.
(Talking about local wisdom, have you feel the crowd of **Ramadan market**? It's very unique.)

Lah Bapak, ramai bana. Beda bana jo suasana **Turun Mandi**.
(Yes I am. It's very different with the atmosphere of **Turun Mandi**.)

Figure 1: Example dialogue between two individuals, with personas incorporated to ensure the conversation reflects their distinct characteristics.

1 Introduction

Dialogue systems have made significant strides in enabling real-life interactions, from task-oriented models that assist users with specific goals, such as booking flights or restaurants (Budzianowski et al., 2018; Chakraborty et al., 2025) or managing schedules (Mo et al., 2024), to chit-chat systems designed for more casual, extended conversations (Lin et al., 2021; Sun et al., 2021). Although a wide range of datasets exist, particularly for open-domain dialogue, most were not created with cultural sensitivity in mind and therefore fail to capture the nuanced ways in which culture shapes human communication. While large language models (LLMs) have substantially advanced the development of dialogue systems, their direct use often makes them struggle to accurately reflect cultural values, particularly when generating culture-specific references or contextually grounded entities (Adilazuarda et al., 2024; Chiu et al., 2024). This limitation highlights the need for culturally enriched dialogue datasets that have contextual relevance in real-world applications.

Previous work has demonstrated that incorporating local entities and leveraging multilingual data augmentation can significantly enhance performance and improve the cultural awareness of LLMs (Ding et al., 2022). Additionally, modeling user personas has been shown to increase the naturalness and engagement of dialogue systems, enabling more personalized and human-like interactions (Zhang et al., 2018). In the context of Southeast Asia, home to over 700 million people across 11 countries,[1], only a limited number of languages have been explored in dialogue system research. Most existing work has focused on languages such as Indonesian (Lin et al., 2021; Kautsar et al., 2023), Thai (Robloke and Kijsirikul,

[1] https://www.worldometers.info/world-population/south-eastern-asia-population/

2019), and Vietnamese (Van et al., 2022), yet key challenges remain due to the region's linguistic diversity and deep cultural influences (Aji et al., 2022). Foremost among these challenges are the scarcity of large-scale annotated dialogue datasets and an overreliance on translated English corpora. Such translations often produce unnatural conversational flows that fail to capture the cultural and linguistic nuances of the target languages, even when local entities are included. Additionally, research on synthetic versus culturally-grounded multi-turn dialogue generation using LLMs, particularly in the context of the Global South, also remains largely overlooked. Existing datasets (Zhang et al., 2018; Ding et al., 2022, 2023) continue to focus predominantly on the Global North and often fail to capture the cultural nuances present in both human-human and human-machine interactions.

To address the aforementioned challenges, we propose SEADIALOGUES, a benchmark dataset featuring multi-turn, culturally grounded, and persona-rich conversations in 8 languages across 6 Southeast Asian countries: Indonesian, Javanese, Minangkabau, Thai, Malay, Vietnamese, Tamil, and Tagalog. It consists of 32,000 dialogues covering over 100 culturally relevant topics, with each dialogue addressing multiple topics, as shown in Figure 1. It also includes 210 diverse personas to support personalized and culturally aware dialogue generation. Our detailed dataset statistics are explained in Appendix A.

Our contributions can be summarized in the following aspects:

- We introduce SEADIALOGUES, a new open-source, multilingual, multi-turn, and persona-rich synthetic dialogue dataset encompassing eight languages across six Southeast Asian countries[2]. The dialogues are LLM-generated and are carefully tailored to reflect local cultures, values, and region-specific topics. Additionally, we construct culturally grounded personas from each country to ensure realistic and contextually accurate interactions.

- We study the effectiveness of generating synthetic multi-turn SEA dialogues with open-weights and proprietary LLMs, assessing their ability to produce culturally appropriate and persona-consistent dialogue.

- We study the correlation between human annotations and LLM judges across different aspects and metrics, finding that G-Eval with the GPT-4.1 mini model exhibits a good correlation with human annotations. However, LLM judges still require improvement to match human evaluations for SEA dialogues.

2 What Factors Make a Good Dataset?

Culturally-Relevant Information or Entities. Translation-based dialogue datasets often directly translate English named entities (e.g., hotels, locations), even when such entities are culturally irrelevant or nonexistent in the target regions, leading to unnatural and impractical interactions (Ding et al., 2022). For instance, a system targeting users in Dubai may inappropriately reference Cambridge-specific entities or postcode systems. Hu et al. (2023) attempt to mitigate this through cultural adaptation strategies such as entity replacement and value redistribution across cities like Dubai, Paris, and Ankara. However, these methods often overlook deeper cultural nuances, regional language variation, and communication styles. As dialogue systems are increasingly deployed in diverse sociocultural settings, grounding them in culturally relevant knowledge is essential for generating coherent, relatable, and effective conversations.

Is LLM enough to generate good data? LLMs have advanced the generation of high-quality synthetic data, addressing challenges related to data scarcity and privacy. Their capacity to produce contextually rich, human-like text supports the creation of training datasets across diverse domains, including healthcare (Peng et al., 2023) and education (Moore et al., 2023). Recent efforts in dialogue data generation leverage different LLMs and few-shot examples from datasets like DialogSum and SAMSum to synthesize dialogue data (Suresh et al., 2025). However, the generation process can suffer from limited domain knowledge, particularly on topics underrepresented in the models' pretraining data. This issue becomes especially pronounced when generating personas and subtopics using zero-shot prompting.

While LLMs implicitly encode a vast amount of cultural knowledge, prior studies have shown that they often fail to apply this knowledge effectively in context. Explicitly providing relevant cultural information has been shown to significantly en-

[2] Upon acceptance, we plan to release the dataset with a CC-BY license.

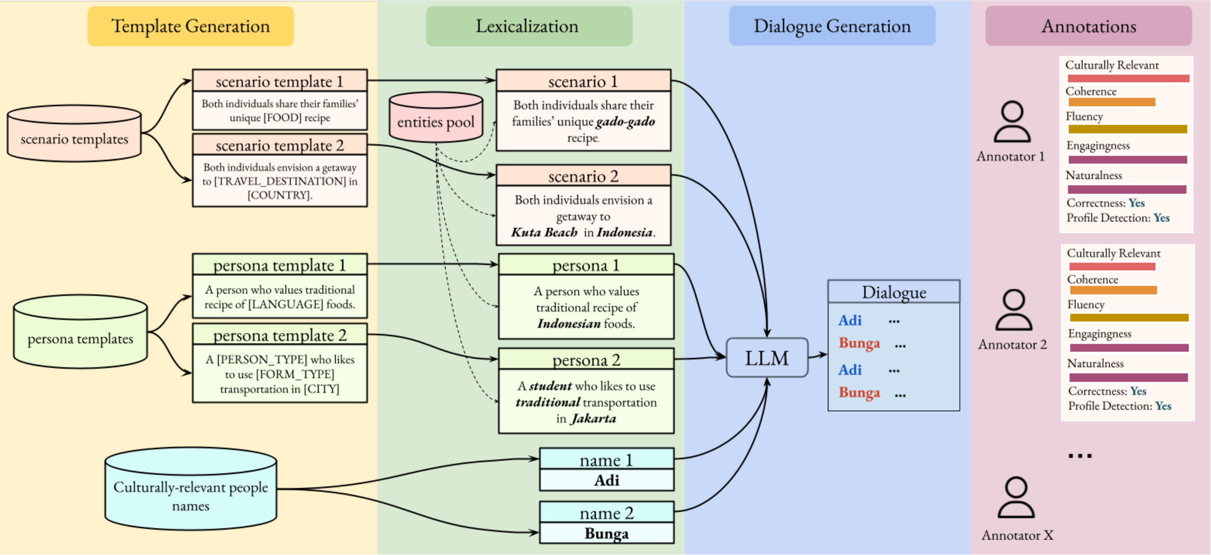

Template Generation
Lexicalization
Dialogue Generation
Annotations
scenario templates
scenario template 1
Both individuals share their families' unique [FOOD] recipe
scenario template 2
Both individuals envision a getaway to [TRAVEL_DESTINATION] in [COUNTRY].
entities pool
scenario 1
Both individuals share their families' unique gado-gado recipe.
scenario 2
Both individuals envision a getaway to Kuta Beach in Indonesia.
persona templates
persona template 1
A person who values traditional recipe of [LANGUAGE] foods.
persona template 2
A [PERSON_TYPE] who likes to use [FORM_TYPE] transportation in [CITY]
persona 1
A person who values traditional recipe of Indonesian foods.
persona 2
A student who likes to use traditional transportation in Jakarta
Culturally-relevant people names
name 1
Adi
name 2
Bunga
LLM
Dialogue
Adi ...
Bunga ...
Adi ...
Bunga ...
Annotator 1
Annotator 2
Annotator X
Culturally Relevant
Coherence
Fluency
Engagingness
Naturalness
Correctness: Yes
Profile Detection: Yes
Culturally Relevant
Coherence
Fluency
Engagingness
Naturalness
Correctness: Yes
Profile Detection: Yes
...

Figure 2: Overview of the SEADIALOGUES generation pipeline, which comprises four key stages: (1) template generation, (2) lexicalization of cultural elements, (3) synthetic dialogue generation using LLMs, and (4) final annotation by human annotators. In the first stage, templates with delexicalized entities are created; these placeholders are then populated with culturally relevance entities during lexicalization. Next, multi-turn dialogues are generated synthetically based on speaker personas. Finally, human annotators evaluate the dialogues using quality metrics.

hance the specificity, cultural sensitivity, and overall quality of responses in intercultural dialogue tasks (Nguyen et al., 2024). These findings suggest that even advanced models benefit from structured and cultural knowledge during data generation.

How does our dataset differ from existing datasets? To address the limitations of prior translation-based and LLM-generated dialogue datasets in representing cultural and local relevance, we propose a multilingual synthetic dialogue dataset covering eight languages across six Southeast Asian countries. Unlike existing approaches that rely solely on zero-shot prompting or entity replacement, our dataset integrates manually curated cultural knowledge, such as local entities, food, and communication norms, directly into the prompt design. Using LLMs, we generate dialogues conditioned on user personas, dialogue topics, and region-specific cultural contexts, aiming to produce coherent, culturally grounded conversations that reflect real-world user behavior. This approach bridges the gap between linguistic diversity and cultural representation in dialogue systems, supporting more inclusive and contextually aware conversational agents for underrepresented regions.

In addition to human evaluation, we propose the use of LLMs as automated judges to assess multi-turn dialogues along dimensions such as coherence, fluency, and cultural relevance. To our knowledge, this is the first dialogue dataset evaluated by LLMs for multi-turn conversations in a multilingual, culturally grounded setting. This dual evaluation framework enhances scalability while maintaining rigor, supporting the development of more inclusive and context-aware dialogue systems for underrepresented regions.

3 SEADIALOGUES

SEADIALOGUES is a multi-turn, multilingual dialogue dataset encompassing eight Southeast Asian languages (Indonesian, Javanese, Malay, Minangkabau, Tagalog, Tamil, Thai, Vietnamese) from six different countries. For comparison, Table 1 highlights how our dataset differs from existing dialogue datasets, with ours being the first to explicitly represent cultural aspects within each conversation.

Figure 2 illustrates the full data construction pipeline. The process begins with the collection of supporting resources, including scenario and persona templates, along with culturally relevant Southeast Asian names. For each dialogue, two domain-relevant scenarios and corresponding personas are selected to ensure consistency and coherence across both intra-scenario and inter-scenario–persona relationships. In the next step, lexicalization is performed by manually curating and inserting matched entities into both scenarios and personas. This step ensures that subtopics are contextually aligned with the personas and maintain linguistic coherence. Once the templates are

Dataset	#Lang.	#Dial.	Topic Type	#Scenario	#Persona	Human Annotations	¬Translation	Cultural Rel.
DailyDialog (Li et al., 2017)	1	13.1K	Single	10	0	✓	✓	×
MultiWOZ (Budzianowski et al., 2018)	1	8.4K	Single	-	-	✓	×	×
PERSONA-CHAT (Zhang et al., 2018)	1	10.9K	Multiple	-	1,155	✓	✓	×
PersonalDialog (Zheng et al., 2019)	1	20.83M	Single	-	-	✓	✓	×
CrossWOZ (Zhu et al., 2020)	1	5K	Single	-	-	✓	✓	×
MuTual (Cui et al., 2020)	1	8.8K	Single	-	-	✓	✓	×
XPersona (Lin et al., 2021)	6	104.6K	Multiple	-	1,155	✓	×	×
GlobalWOZ (Ding et al., 2022)	21	9.4K	Single	-	-	✓	×	×
Multi2WOZ (Hung et al., 2022)	5	1K	Single	-	-	✓	×	×
Multi3WOZ (Hu et al., 2023)	4	8.2K	Single	-	-	✓	✓	×
XDailyDialog (Liu et al., 2023b)	4	52K	Single	10	0	✓	×	×
SEADialogues	**8**	**32K**	**Multiple**	**300**	**210**	✓	✓	✓

Table 1: Comparison of dialogue dataset statistics. Our dataset is the only dataset in the list that focuses on generating culture-related entities in the multi-turn conversational dataset. #Dial. represents the number of dialogues, and ¬Translation resembles whether the dataset is from the translation of an existing dataset.

finalized, dialogue generation is carried out by prompting LLMs with carefully crafted instructions that incorporate the lexicalized scenarios, personas, and the selected cultural names.

The generated dialogues are subsequently evaluated through a two-fold approach: human annotation for qualitative assessment and automatic evaluation using G-Eval (Liu et al., 2023a), M-Prometheus (Pombal et al., 2025), and R3 (Anugraha et al., 2025) to provide quantitative insights. This structured pipeline ensures the generation of high-quality, culturally appropriate dialogue data for our main objective in this study.

3.1 Template Generation

To craft rich, multi-turn dialogue data, we begin by curating several reusable resources we call templates. Mainly, there are two templates to construct the dialogue setup, which are scenario and speakers' persona templates. Both are derived from our curated topics list, which consists of 100 topics in total. To build them, we employ the GPT-4.1 mini (Achiam et al., 2023) LLM model to generate 300 scenarios, where 1 topic corresponds to 3 scenarios. For each scenario, we identify culturally grounded entities and replace them with abstract placeholders. These placeholder slots will later be instantiated with values drawn from a curated pool of culturally relevant entities. This approach allows the same dialogue scenario to be adapted across multiple country contexts while maintaining cultural fidelity. After all scenario templates have been generated, we conduct human annotation to identify and revise low-quality or inappropriate templates. See Table 6 for the scenario template examples.

In parallel, we also develop persona templates to characterize the personality traits of each speaker. These templates guide the linguistic expression and behavioral tendencies throughout the dialogue. Similar to topic templates, any culture-specific references within the persona descriptions are masked and later filled with values from a corresponding cultural entity pool. Initially, we aimed to have the same number of personas as scenario templates, which is 300 in total. However, we decided to drop 90 personas due to the low quality of their associated generated templates. See Table 7 for the persona template examples.

Each dialogue setup consists of two scenarios. This multi-scenario design mirrors the dynamic nature of real-world conversations, which often shift fluidly between different topics rather than remaining fixed on a single subject. To address this, at the end of this step, we select two scenario templates and two persona templates (per participant in the two-way exchange) for use in the subsequent lexicalization stage for each generated dialogue.

3.2 Lexicalization

Lexicalization is essential for embedding cultural nuances into dialogue. To achieve this, we incorporate entity lists tagged with language codes, such as `-ind` for Indonesian, `-tha` for Thai, or the generic (entities that can be used in all languages) `-gen`, to indicate their intended scope of use. Examples of these tagged entities are provided in Table 8.

Those entities are used to generating specific scenarios by lexicalizing the scenario templates. Using the selected templates from the previous step, each slot is systematically filled with entities whose language tags match the target language of the dialogue. To ensure comprehensive coverage, all valid combinations generated through this slot-filling procedure are enumerated. For example, to generate Indonesian dialogue, the entity 'iconic ricepaddies of Ubud-ind' can be inserted into the template 'Person A describes a family trip to the [TRAVEL_DESTINATION]' to produce: "Person A describes a family trip to the iconic rice paddies of Ubud."

Concurrently, speaker personas are generated to populate the dialogues. The slots within this template are then filled with appropriate entities from the compiled lists. To add depth and individuality to each persona, one personality trait is randomly selected from the list of personality traits and appended to the description, thereby enriching the character profile.

Preserve and validate contextual dependencies. To ensure the dialogues remain authentic and culturally relevant, contextual alignment between certain slots within a template must be retained. We build a dictionary in JSON files for pairs of slots that are inherently linked either culturally or semantically. It helps exclude combinations that would be incompatible or nonsensical, thereby maintaining the coherence and realism of the dialogue. For instance, if there are both `[COUNTRY]` and `[CITY]` entities within a template, for `[CITY]` values like Jakarta, Bandung, and Denpasar will only be paired with Indonesia as `[COUNTRY]` value, or Bangkok, Chiang Mai, and Songkhla will only be paired with Thailand.

3.3 Dialogue Generation

Once all necessary components are collected, including the target language for the dialogue, lexicalized scenarios, lexicalized personas, personalities, and character names, we proceed to the dialogue generation phase. In this phase, we curate a prompt designed to guide the model in generating culturally appropriate dialogue and input it into both open-source and proprietary LLMs. The prompt used for dialogue generation is shown in Appendix B. Additionally, we specify a maximum number of dialogue turns within the prompt to prevent overly long or unnatural conversations.

3.4 Annotations

To ensure the quality and validity of the dataset, we employ a structured human annotation process. Each dialogue is assessed based on the following criteria to evaluate conversational abilities: (1) *Fluency*, (2) *Engagingness*, (3) *Coherence*, (4) *Naturalness*, and (5) *Culturally Relevance*. In addition, annotators conduct these evaluations to measure instruction-following abilities: (1) *Profile Detection* and (2) *Correctness*. Detailed annotation scoring rubrics and guidelines are provided in Appendices D and E. For each language, three annotators are employed to perform annotations on our platform. They were hired through our contacts and are indigenous people from the country, fluent in the native language. To see an overview of our platform, please refer to Appendix F.

3.5 Automatic Evaluation

Relying solely on human evaluation poses several challenges, including significant time requirements, logistical complexities, and inherent subjectivity. While human evaluation is our primary measure of dialogue quality, we strategically incorporate automatic evaluation methods. The main goal of using these automatic evaluation methods is to ensure the scalability of our data creation process.

We utilize G-EVAL (Liu et al., 2023a), a prompt-based evaluator as the LLM-as-judge. The prompt provides information regarding the definition of the evaluation task and the assessment criteria. It also employs a chain of thought, consisting of a series of instructions that outline the evaluation steps. Additionally, it includes a scoring function that interacts with a language model using a form-filling approach and probabilities of the return tokens. In addition to G-EVAL, we also use the R3 (Anugraha et al., 2025) and M-Prometheus (Pombal et al., 2025) reward model as the LLM-as-judge. These methods use point-wise evaluation, which assesses the quality of a single response by assigning an integer score based on specific criteria. The reward model takes task instructions, responses, and rubrics as input, generates the scores, and provides explanations with its reasoning model capability. For the details of instructions, rubrics, and prompts, see section G in the Appendix.

4 Experimental Setup

4.1 Dialogue Generation

We employ four different models for dialogue generation, which include both closed-source and open-source options. For our open-source models, we use Llama-3.1-8B Instruct (Grattafiori et al., 2024) and Aya-8B Expanse (Li et al., 2017). For the closed-source models, we utilize Gemini-Flash-1.5 (Team et al., 2024) and GPT-4o mini (Achiam et al., 2023). We run these models on an Nvidia A100 40GB and use HuggingFace, OpenAI, and

Model	ind	jav	min	tam	tgl	tha	vie	zsm
Coherence								
Aya-8B-Expanse	**3.00 ± 0.00**	**3.00 ± 0.00**	3.00 ± 0.01	2.42 ± 0.55	2.34 ± 0.62	2.41 ± 0.50	**3.00 ± 0.00**	**3.00 ± 0.00**
Gemini 1.5 Flash	**3.00 ± 0.00**	**3.00 ± 0.00**	**3.00 ± 0.00**	**3.00 ± 0.00**	**3.00 ± 0.00**	**3.00 ± 0.00**	**3.00 ± 0.00**	3.00 ± 0.01
GPT-4o mini	**3.00 ± 0.00**	**3.00 ± 0.00**	**3.00 ± 0.00**	**3.00 ± 0.00**	**3.00 ± 0.00**	**3.00 ± 0.00**	**3.00 ± 0.00**	**3.00 ± 0.00**
Llama-3.1-Instruct	3.00 ± 0.01	1.97 ± 0.76	2.59 ± 0.66	2.58 ± 0.44	2.57 ± 0.60	2.90 ± 0.32	2.98 ± 0.10	2.99 ± 0.06
Culturally Relevance								
Aya-8B-Expanse	2.98 ± 0.10	2.98 ± 0.15	**2.99 ± 0.04**	2.55 ± 2.67	2.92 ± 5.49	1.33 ± 0.81	3.00 ± 0.01	2.98 ± 0.10
Gemini 1.5 Flash	**3.00 ± 0.03**	**3.00 ± 0.00**	2.99 ± 0.08	**3.00 ± 0.01**	**3.00 ± 0.02**	**2.99 ± 0.04**	3.00 ± 0.01	**3.00 ± 0.03**
GPT-4o mini	2.99 ± 0.03	3.00 ± 0.03	2.97 ± 0.10	3.00 ± 0.02	2.99 ± 0.03	2.99 ± 0.06	**3.00 ± 0.00**	2.99 ± 0.06
Llama-3.1-Instruct	2.95 ± 0.20	2.09 ± 2.10	2.49 ± 0.82	2.25 ± 0.93	2.66 ± 0.53	2.79 ± 0.42	2.92 ± 0.23	2.89 ± 0.33
Engagingness								
Aya-8B-Expanse	**2.59 ± 0.36**	**2.62 ± 0.34**	**2.64 ± 0.34**	1.70 ± 0.39	1.71 ± 0.39	1.55 ± 0.32	**2.66 ± 0.36**	**2.56 ± 0.37**
Gemini 1.5 Flash	2.30 ± 0.35	2.31 ± 0.38	2.21 ± 0.36	2.04 ± 0.43	2.08 ± 0.33	2.42 ± 0.43	2.38 ± 0.38	2.26 ± 0.39
GPT-4o mini	2.42 ± 0.35	2.44 ± 0.35	2.35 ± 0.36	**2.14 ± 0.31**	**2.57 ± 0.34**	**2.45 ± 0.37**	2.47 ± 0.39	2.38 ± 0.38
Llama-3.1-Instruct	1.86 ± 0.26	1.17 ± 0.33	1.52 ± 0.45	1.26 ± 0.24	1.32 ± 0.34	1.58 ± 0.41	1.82 ± 0.39	1.84 ± 0.33
Fluency								
Aya-8B-Expanse	3.00 ± 0.00	**3.00 ± 0.00**	**3.00 ± 0.00**	2.29 ± 0.50	1.45 ± 0.49	1.56 ± 0.48	**3.00 ± 0.00**	**3.00 ± 0.00**
Gemini 1.5 Flash	3.00 ± 0.01	**3.00 ± 0.00**	2.99 ± 0.02	**3.00 ± 0.00**	**3.00 ± 0.00**	**3.00 ± 0.04**	**3.00 ± 0.00**	**3.00 ± 0.00**
GPT-4o mini	3.00 ± 0.00	**3.00 ± 0.00**	**3.00 ± 0.00**	2.99 ± 0.06	**3.00 ± 0.00**	2.99 ± 0.09	**3.00 ± 0.00**	**3.00 ± 0.00**
Llama-3.1-Instruct	3.00 ± 0.01	1.60 ± 0.71	2.43 ± 0.75	2.54 ± 0.44	2.31 ± 0.66	2.71 ± 0.51	2.92 ± 0.25	2.97 ± 0.15
Naturalness								
Aya-8B-Expanse	3.00 ± 0.03	**3.00 ± 0.00**	**3.00 ± 0.00**	1.98 ± 0.45	1.32 ± 0.35	1.21 ± 0.25	**3.00 ± 0.00**	**3.00 ± 0.00**
Gemini 1.5 Flash	2.98 ± 0.10	2.97 ± 0.08	2.99 ± 0.05	2.97 ± 0.12	2.91 ± 0.20	**2.98 ± 0.08**	2.99 ± 0.04	2.98 ± 0.07
GPT-4o mini	**3.00 ± 0.01**	2.98 ± 0.12	3.00 ± 0.02	**2.97 ± 0.10**	**3.00 ± 0.02**	**2.98 ± 0.08**	**3.00 ± 0.00**	3.00 ± 0.01
Llama-3.1-Instruct	2.60 ± 0.46	1.17 ± 0.43	1.66 ± 0.74	1.69 ± 0.44	1.43 ± 0.52	1.77 ± 0.56	2.23 ± 0.69	2.33 ± 0.56
Correctness								
Aya-8B-Expanse	0.98 ± 0.10	0.99 ± 0.04	1.00 ± 0.02	0.44 ± 0.38	0.66 ± 0.34	0.48 ± 0.39	**1.00 ± 0.00**	0.98 ± 0.10
Gemini 1.5 Flash	**1.00 ± 0.00**	**1.00 ± 0.01**	**1.00 ± 0.00**	**1.00 ± 0.00**	1.00 ± 0.01	**1.00 ± 0.00**	1.00 ± 0.01	**1.00 ± 0.00**
GPT-4o mini	1.00 ± 0.01	1.00 ± 0.02	1.00 ± 0.02	1.00 ± 0.01	**1.00 ± 0.00**	**1.00 ± 0.00**	0.99 ± 0.06	**1.00 ± 0.00**
Llama-3.1-Instruct	0.98 ± 0.11	0.77 ± 0.35	0.91 ± 0.26	0.41 ± 0.36	0.96 ± 0.13	0.97 ± 0.13	0.99 ± 0.06	0.98 ± 0.09
Profile Detection								
Aya-8B-Expanse	0.96 ± 0.13	0.98 ± 0.06	0.97 ± 0.09	0.70 ± 0.27	0.63 ± 0.30	0.51 ± 0.27	**0.99 ± 0.06**	0.96 ± 0.11
Gemini 1.5 Flash	**0.97 ± 0.10**	0.98 ± 0.06	**0.99 ± 0.04**	**0.99 ± 0.04**	0.95 ± 0.13	0.98 ± 0.06	0.98 ± 0.08	0.95 ± 0.14
GPT-4o mini	0.96 ± 0.14	**0.99 ± 0.03**	0.96 ± 0.12	0.97 ± 0.09	0.97 ± 0.09	0.96 ± 0.12	0.98 ± 0.08	0.98 ± 0.07
Llama-3.1-Instruct	0.93 ± 0.17	0.59 ± 0.35	0.80 ± 0.29	0.70 ± 0.28	0.78 ± 0.25	0.83 ± 0.23	**0.97 ± 0.07**	0.94 ± 0.11

Table 2: G-Eval results on conversational generation and instruction following capabilities. Coherence, Engagingness, Fluency, and Naturalness have a score range of 1 to 3, while Culturally Relevance has a score range of 0 to 3. For instruction-following metrics, Profile Detection and Correctness have a binary score.

Google API as packages to run these models. In line with best practices observed in prior work on dialogue generation using LLMs (Li et al., 2024; Ye et al., 2024), we search for the best hyperparameter by comparing the dialogue, and set the sampling parameters to a temperature (T) of 0.7 and a (*top-P*) of 0.8. At the end of the pipeline, we successfully collected a total of 32,000 dialogues, distributed evenly as 4,000 dialogues per language, with each model contributing a unified set of 1,000 dialogues under this generation scheme.

4.2 Automatic Evaluation

We utilize automatic evaluation through LLM-as-judge methods, including G-Eval (Liu et al., 2023a) as our primary method, M-Prometheus (Pombal et al., 2025), and R3 reward model (Anugraha et al., 2025) in our dataset. For G-Eval, we use the GPT-4.1 mini model; for M-Prometheus, we apply the M-Prometheus-7B model; and for R3 reward model, we utilize the R3-Qwen-14B-14k version. These automatic methods assess the generated dialogues using the same metrics that we applied during our human annotation process.

5 Results and Analysis

5.1 Human Evaluation

As shown in Table 3, in terms of conversational capabilities, both GPT-4o Mini and Gemini 1.5 Flash demonstrate the best performance across various metrics. However, Gemini 1.5 Flash leads on most metrics, particularly in fluency for the Minangkabau and Thai languages. Among the open-source models, Aya-8B-Expanse performs well with Javanese and Minangkabau, though it falls short on the fluency metric. In contrast, Llama-3.1 Instruct generally produces the lowest scores across most metrics.

Model	ind	jav	min	tha
Coherence				
Aya-8B-Expanse	2.79 ± 0.44	2.84 ± 0.43	2.81 ± 0.40	1.98 ± 0.73
Gemini 1.5 Flash	**2.88 ± 0.34**	2.90 ± 0.34	**2.82 ± 0.40**	**2.46 ± 0.70**
GPT-4o mini	2.84 ± 0.37	**2.92 ± 0.30**	2.75 ± 0.45	2.45 ± 0.64
Llama-3.1-Instruct	2.70 ± 0.49	1.17 ± 0.48	2.41 ± 0.73	2.19 ± 0.75
Culturally Relevance				
Aya-8B-Expanse	2.02 ± 0.98	1.93 ± 1.29	2.15 ± 1.04	0.71 ± 0.93
Gemini 1.5 Flash	**2.30 ± 0.86**	2.22 ± 1.19	**2.17 ± 0.98**	**1.59 ± 1.15**
GPT-4o mini	2.11 ± 0.92	**2.23 ± 1.18**	2.16 ± 1.04	1.42 ± 1.15
Llama-3.1-Instruct	1.95 ± 0.88	0.36 ± 0.80	1.88 ± 1.16	1.37 ± 1.12
Engagingness				
Aya-8B-Expanse	**2.62 ± 0.52**	**2.75 ± 0.47**	**2.72 ± 0.45**	1.61 ± 0.59
Gemini 1.5 Flash	2.51 ± 0.52	2.56 ± 0.54	2.62 ± 0.49	**2.26 ± 0.70**
GPT-4o mini	2.35 ± 0.50	2.68 ± 0.49	2.58 ± 0.52	2.04 ± 0.72
Llama-3.1-Instruct	2.10 ± 0.63	1.21 ± 0.55	2.30 ± 0.71	1.93 ± 0.60
Fluency				
Aya-8B-Expanse	2.75 ± 0.47	1.09 ± 0.41	1.48 ± 0.73	1.70 ± 0.58
Gemini 1.5 Flash	**2.92 ± 0.28**	2.72 ± 0.48	**2.32 ± 0.54**	**2.30 ± 0.60**
GPT-4o mini	2.88 ± 0.32	**2.73 ± 0.47**	1.46 ± 0.68	2.15 ± 0.61
Llama-3.1-Instruct	2.83 ± 0.39	1.04 ± 0.22	1.50 ± 0.59	1.98 ± 0.57
Naturalness				
Aya-8B-Expanse	2.49 ± 0.57	1.46 ± 0.69	2.34 ± 0.52	1.35 ± 0.48
Gemini 1.5 Flash	**2.60 ± 0.52**	**2.30 ± 0.76**	**2.46 ± 0.53**	**2.15 ± 0.53**
GPT-4o mini	2.28 ± 0.47	2.26 ± 0.75	2.24 ± 0.55	1.83 ± 0.50
Llama-3.1-Instruct	2.09 ± 0.48	1.02 ± 0.14	1.91 ± 0.72	1.64 ± 0.48
Correctness				
Aya-8B-Expanse	0.91 ± 0.28	0.97 ± 0.17	0.99 ± 0.08	0.65 ± 0.48
Gemini 1.5 Flash	**0.98 ± 0.15**	0.96 ± 0.19	**1.00 ± 0.06**	0.88 ± 0.32
GPT-4o mini	**0.98 ± 0.15**	**0.98 ± 0.13**	0.99 ± 0.08	**0.91 ± 0.29**
Llama-3.1-Instruct	0.90 ± 0.30	0.25 ± 0.43	0.86 ± 0.34	0.87 ± 0.33
Profile Detection				
Aya-8B-Expanse	**0.65 ± 0.48**	0.81 ± 0.39	**0.95 ± 0.22**	0.40 ± 0.49
Gemini 1.5 Flash	0.57 ± 0.49	0.80 ± 0.40	0.92 ± 0.27	**0.81 ± 0.39**
GPT-4o mini	0.59 ± 0.49	**0.83 ± 0.38**	0.94 ± 0.25	0.77 ± 0.42
Llama-3.1-Instruct	0.59 ± 0.49	0.19 ± 0.39	0.81 ± 0.40	0.71 ± 0.45

Table 3: Human Annotations Results.

For the following instructions abilities, Gemini 1.5 Flash and GPT-4o mini exhibit good performance, nearly 100 percent accuracy on Correctness. However, they have lower performance on Profile Detection metrics. For Aya-8B-Expanse, it shows similar performance with Gemini 1.5 Flash and GPT-4o mini for Indonesian, Javanese, and Minangkabau, but struggles with Thai. Lastly, in the case of Llama-3.1 Instruct, its overall performance is generally lower, except when compared to Aya-8B-Expanse in Thai.

5.2 Automatic Evaluation

In terms of conversational quality, as shown in Table 2, closed-weight models like Gemini 1.5 Flash and GPT-4o mini generally achieve higher scores across most metrics and languages. In contrast, open-weight models, particularly Llama-3.1-Instruct, display more variability and show lower performance. On a positive note, the Aya-8B-Expanse model receives comparably high scores for languages such as Indonesian, Javanese, Minangkabau, Vietnamese, and Malay, ranking among the best for engagingness and naturalness in these languages.

	G-Eval	**M-Prometheus**	**R3**
Pearson			
Coherence	**0.5876**	0.4734	0.5227
Culturally Relevance	**0.4832**	0.4324	0.3027
Engagingness	0.5401	**0.5499**	0.4627
Fluency	0.1528	0.1749	**0.1949**
Naturalness	**0.6124**	0.5228	0.4501
Spearman			
Coherence	**0.4408**	0.3088	0.3994
Culturally Relevance	**0.2506**	0.2241	0.2146
Engagingness	**0.4486**	0.4462	0.3754
Fluency	0.0372	0.1346	**0.1429**
Naturalness	**0.5307**	0.4650	0.3830
Kendall Tau			
Coherence	**0.4100**	0.2893	0.3765
Culturally Relevance	**0.2234**	0.2069	0.1969
Engagingness	0.3588	**0.4098**	0.3482
Fluency	0.0326	0.1220	**0.1300**
Naturalness	**0.4660**	0.4270	0.3551

Table 4: Automatic Evaluations and Human Annotations Correlations on Minangkabau.

For instruction-following to generate the dialogue capabilities, Gemini 1.5 Flash and GPT-4o mini demonstrate strong results in both correctness and profile detection. Llama-3.1-Instruct performs adequately in correctness for specific languages, although it struggles with Tamil and Javanese. However, it shows poorer results in profile detection for most languages. Similar to its conversational capabilities, Aya-8B-Expense shows promising results in both correctness and profile detection for Indonesian, Javanese, Minangkabau, Vietnamese, and Malay, but performs less effectively with other languages. These results indicate similar trends in both Human Evaluation and Automatic Evaluation, particularly in pointing out which model can generate good dialogue for specific languages. See the result details in Appendix G.

5.3 Human Annotation Alignment

Figure 3 presents a visual analysis of the correlation between the automatic evaluation scores and human judgments for key conversational quality metrics, such as fluency, coherence, engagingness, and naturalness. All metrics show positive correlations between automatic and human evaluations. G-Eval aligns more closely with human judgment than other automatic evaluation methods, as shown in Table 4. The most significant challenges lie in the Fluency and Cultural Relevance metrics, as evidenced by their lower correlation compared to other metrics.

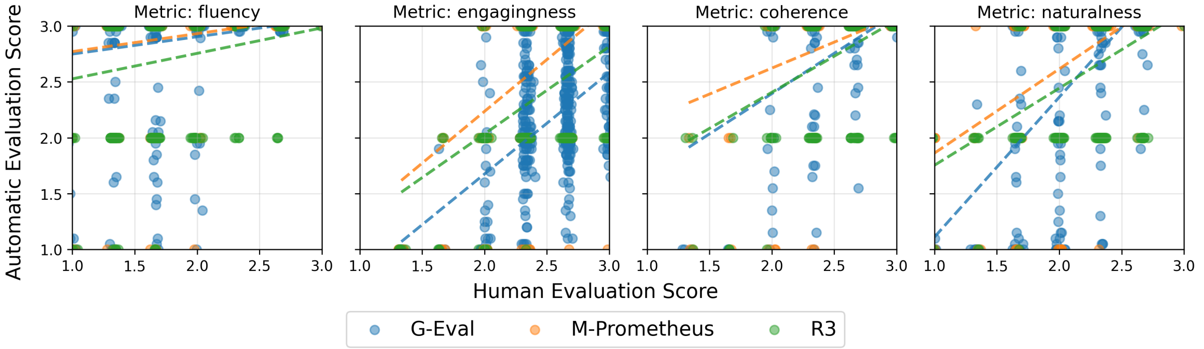


Figure 3: Correlation between Automatic Evaluations and Human Annotations per Metric on Minangkabau

6 Related Work

6.1 Personalized Conversations

Dialogue research has primarily focused on task-oriented systems (Budzianowski et al., 2018; Ding et al., 2022; Goel et al., 2023; He et al., 2024) and question-answering dialogues (Feng et al., 2020; Rajpurkar et al., 2016), but these approaches often lack the richness of open-domain, human-human interactions. Datasets like DailyDialog (Li et al., 2017) and XDailyDialog (Liu et al., 2023b) address this by offering multi-turn, intent- and emotion-annotated dialogues for more naturalistic chit-chat.

Building on this, persona-conditioned datasets such as PERSONA-CHAT (Zhang et al., 2018) and PersonalDialog (Zheng et al., 2019) simulate personalized conversations using user profiles. Recent work like BotChat (Duan et al., 2023) uses LLMs to generate scalable, persona-driven dialogues from seed prompts. Multilingual datasets, including XPersona (Lin et al., 2021) and XDailyDialog, improve cross-lingual transfer through translated and refined dialogues. While some datasets constrain dialogue to a single topic, others like PERSONA-CHAT and BotChat allow topic shifts, better mirroring real-world conversation dynamics. Finally, cultural grounding has gained importance. Efforts like GlobalWOZ (Ding et al., 2022) localize templates to reflect cultural norms—a principle we also adopt in our dataset design.

6.2 Dataset Evaluation

Recent work increasingly adopts LLM-based evaluation to complement human assessment. BotChat (Duan et al., 2023) introduces a three-part framework: UniEval for single-model judgments, BotChat Arena for pairwise comparisons, and G-Eval (Liu et al., 2023a) for aligning model outputs with human references. XDailyDialog (Liu et al., 2023b) combines automatic metrics (e.g., BLEU, F1, DIST-n) with human evaluations at both turn and dialogue levels, assessing fluency, relevance, and coherence. Inspired by these practices, we adopt a hybrid evaluation protocol that integrates human ratings with LLM-based scoring using G-Eval, enabling scalable and consistent quality assessment.

7 Conclusion

We introduce SEADIALOGUES, an open-source, multilingual, multi-turn, and persona-rich synthetic dialogue dataset that spans eight languages across six Southeast Asian countries. Motivated by the impressive generative capabilities of large language models, we incorporates structured guardrails in the data generation pipeline, including scenario and persona templates as well as culturally grounded lexicalization strategies. We perform further study on samples from our dataset, by conducting a comprehensive evaluation using both human annotation and LLM-as-a-judge assessments across several key metrics. These include fluency, coherence, naturalness, engagingness, cultural relevance, persona consistency, and factual correctness. Our findings indicate that proprietary (closed-weight) models generally outperform open-weight models on these dimensions. This highlights a pressing need for high-quality, culturally enriched, and persona-aware datasets like ours to support the development of open-weight LLMs. In turn, such resources can help bridge the quality gap with proprietary models. Additionally, improving the cultural and persona grounding in datasets may enhance the alignment of LLM-as-a-judge systems with human evaluators in future dialogue assessment tasks.

Limitations

This paper focuses on approximating natural human-human conversations using large language models. However, it does not yet include targeted evaluations on specific benchmarks such as topic transition detection (Soni et al., 2021) or persona detection (Jun and Lee, 2025). These tasks are particularly relevant for capturing conversational nuances in Southeast Asian cultural contexts. While this work represents an initial step toward understanding LLM-generated dialogue in Southeast Asian settings, future work should incorporate more comprehensive benchmarking and explore task-specific methodologies to better assess and improve performance in culturally grounded dialogue generation.

Additionally, although humans manually curate the topics, names, and entities, there is still a risk that the generated dialogues may not fully capture the cultural nuances of certain languages, as they are produced using various LLMs. However, we make every effort to ensure cultural appropriateness by carefully curating the seed entities and related content for each region.

Ethics Statement

Throughout our study, we commit to adhering to ethical standards and best practices in NLP research. The dialogue data includes character names selected from manually curated name lists representative of each country. These name pools are created using publicly available, non-sensitive sources (e.g., common baby name registries), and do not reference or target any real individuals. While care has been taken to ensure these are generic, there remains a small possibility that some names may coincidentally match real individuals; any such resemblance is purely coincidental.

To minimize harm, all dialogue topics were manually curated to avoid discussions involving violence or other sensitive content. As a result, we believe the likelihood of harmful or inappropriate material appearing in the dataset is very low. All annotators were compensated fairly, following wage standards in their respective countries. Additionally, all annotators agree for their annotations to be publicly released in an aggregated form (e.g., scores), with no personally identifiable information included. We have taken steps to ensure that annotator privacy and confidentiality are fully maintained.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Muhammad Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, et al. 2022. One country, 700+ languages: Nlp challenges for underrepresented languages and dialects in indonesia. In *60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, pages 7226–7249. Association for Computational Linguistics (ACL).

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

David Anugraha, Zilu Tang, Lester James V. Miranda, Hanyang Zhao, Mohammad Rifqi Farhansyah, Garry Kuwanto, Derry Wijaya, and Genta Indra Winata. 2025. R3: Robust rubric-agnostic reward models.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Amartya Chakraborty, Paresh Dashore, Nadia Bathaee, Anmol Jain, Anirban Das, Shi-Xiong Zhang, Sambit Sahu, Milind Naphade, and Genta Indra Winata. 2025. T1: A tool-oriented conversational dataset for multi-turn agentic planning. *arXiv preprint arXiv:2505.16986*.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. 2024. Culturalbench: A robust, diverse, and challenging cultural benchmark by human-ai culturalteaming. *arXiv preprint arXiv:2410.02677*.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416.

Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022.

Globalwoz: Globalizing multiwoz to develop multilingual task-oriented dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1639–1657.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051.

Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2023. Botchat: Evaluating llms' capabilities of having multi-turn dialogues. *arXiv preprint arXiv:2310.13650*.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128.

Rahul Goel, Waleed Ammar, Aditya Gupta, Siddharth Vashishtha, Motoki Sano, Faiz Surani, Max Chang, HyunJeong Choe, David Greene, Chuan He, et al. 2023. Presto: A multilingual dataset for parsing realistic task-oriented dialogs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10820–10833.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. 2024. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*.

Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023. Multi 3 woz: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems. *Transactions of the Association for Computational Linguistics*, 11:1396–1415.

Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. Multi2woz: A robust multilingual dataset and conversational pretraining for task-oriented dialog. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3687–3703.

Yonghyun Jun and Hwanhee Lee. 2025. Exploring persona sentiment sensitivity in personalized dialogue generation.

Muhammad Kautsar, Rahmah Nurdini, Samuel Cahyawijaya, Genta Indra Winata, and Ayu Purwarianti. 2023. Indotod: A multi-domain indonesian benchmark for end-to-end task-oriented dialogue systems. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 85–99.

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.

Yiwei Li, Fei Mi, Yitong Li, Yasheng Wang, Bin Sun, Shaoxiong Feng, and Kan Li. 2024. Dynamic stochastic decoding strategy for open-domain dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11585–11596, Bangkok, Thailand. Association for Computational Linguistics.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin 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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Zeming Liu, Ping Nie, Jie Cai, Haifeng Wang, Zheng-Yu Niu, Peng Zhang, Mrinmaya Sachan, and Kaiping Peng. 2023b. Xdailydialog: A multilingual parallel dialogue corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12240–12253.

Lingbo Mo, Shun Jiang, Akash Maharaj, Bernard Hishamunda, and Yunyao Li. 2024. Hiertod: A task-oriented dialogue system driven by hierarchical goals. *arXiv preprint arXiv:2411.07152*.

Steven Moore, Richard Tong, Anjali Singh, Zitao Liu, Xiangen Hu, Yu Lu, Joleen Liang, Chen Cao, Hassan Khosravi, Paul Denny, et al. 2023. Empowering education with llms-the next-gen interface and content generation. In *International Conference on Artificial Intelligence in Education*, pages 32–37. Springer.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024. Cultural commonsense knowledge for intercultural dialogues. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1774–1784.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. 2023. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André F. T. Martins. 2025. M-prometheus: A suite of open multilingual llm judges.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Ramon Robloke and Boonserm Kijsirikul. 2019. A task-oriented dialogue bot using long short-term memory with attention for thai language. In *Proceedings of the 1st International Conference on Advanced Information Science and System*, pages 1–6.

Mayank Soni, Brendan Spillane, Emer Gilmartin, Christian Saam, Benjamin R Cowan, and Vincent Wade. 2021. An empirical study of topic transition in dialogue.

Kai Sun, Seungwhan Moon, Paul A Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chat to enhance task-oriented dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1570–1583.

Sathya Krishnan Suresh, Wu Mengjun, Tushar Pranav, and EngSiong Chng. 2025. Diasynth: Synthetic dialogue generation framework for low resource dialogue applications. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 673–690.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Phi Nguyen Van, Tung Cao Hoang, Dung Nguyen Manh, Quan Nguyen Minh, and Long Tran Quoc. 2022. Vi-woz: A multi-domain task-oriented dialogue systems dataset for low-resource language. *arXiv preprint arXiv:2203.07742*.

Guanghui Ye, Huan Zhao, Zixing Zhang, Xupeng Zha, and Zhihua Jiang. 2024. Lstdial: Enhancing dialogue generation via long-and short-term measurement feedback. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5857–5871.

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.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.

A Dataset Statistics

Table 5 shows the detailed statistics of SEADIALOGUES.

No. of languages	8
No. of dialogues	32,000
Average dialogues per language	4,000
Average utterances per dialogue	13.86
Average words per utterance	21.69
No. of topics	100
Topics per dialogue	2
No. of scenarios	300
No. of personas	210

Table 5: SEADialogue statistics.

B Persona and Topics

This section presents examples of documents used for data generation within our framework.

- Table 6 presents samples of scenario templates for dialogue.
- Table 7 provides examples of persona templates.
- Table 8 lists entities use to lexicalize the scenario and persona templates.
- Table 9 showcases various personality traits to complete the persona information.
- Table 10 contains samples of names for the characters in the dialogue.
- Table 11 shows the mapping of coupled entities, specifically when a scenario or persona template includes coupled delexicalized entities.

- Finally, Table 12 displays the prompt utilized for generating dialogue.

C Dialogue Generation Details

Incorporating multiple scenarios into a single dialogue prompt raises the risk of abrupt or unnatural topic transitions. To address this, we employ TOP2VEC (Angelov, 2020), a topic modeling method that clusters our curated pool of scenarios based on semantic similarity. This enables us to select pairs of scenarios from within the same cluster, ensuring thematic coherence and smoother transitions between topics.

D Annotation Rubrics

To assess the quality of generated dialogues, we employ a comprehensive evaluation rubric (see Table 13) consisting of six criteria: Fluency, Engagingness, Coherence, Naturalness, Cultural Relevance, Profile Detection, and Correctness. Each criterion targets a specific aspect of conversational quality, ensuring both linguistic and contextual alignment with the intended design of the dialogue system.

E Human Annotation Guidelines

Each annotation unit consists of a prompt and a multi-turn dialogue generated in response. The prompt includes two topics and two speaker personas. Annotators evaluate the dialogue with respect to quality, persona alignment, and topic relevance through the following steps:

Step 1: Read the Prompt Annotators first read the prompt to identify the two intended **topics** and the two **personas**, which include details such as speaker background, personality traits, and interests. Cultural or linguistic context (e.g., Indonesian, Javanese, Minangkabau, or Thai) should also be noted when applicable.

Step 2: Read the Dialogue Annotators then read the full multi-turn dialogue to understand its tone, structure, and alignment with the prompt. This step ensures that judgments are based on the overall flow and not isolated turns.

Step 3: Score the Dialogue Each dialogue is then rated based on the annotation rubrics as mentioned in Appendix D.

Step 4: Use Examples The guideline includes detailed examples in English, Indonesian, Minangkabau, and Thai for each score level, helping annotators apply the criteria consistently across languages and domains.

F Human Annotation Platform

Figures 4a–4d show the proprietary annotation platform we built to support human annotators in performing their tasks.

G Automatic Evaluation

G.1 G-Eval

We perform G-Eval by modifying the original G-Eval prompt template (Liu et al., 2023a). Table 14 shows the one that we use. Every metric follows this general template, substituting placeholders with its metric information. The *metric evaluation criteria* correspond to the Evaluation Criteria column in Table 13. For *metric evaluation steps*, each metric has its specific procedures, which are shown in Table 15. Additionally, for some metrics, they provide additional information before presenting the dialogue. Language information is added for the *Fluency*, *Naturalness*, and *Culturally Relevance* metrics. Topic descriptions are provided for the *Correctness* metrics. Lastly, Persona descriptions are given for *Profile Detection* metrics.

For M-Prometheus, we run the model using a prompt similar to the one presented in the paper (Pombal et al., 2025). Table 16 shows the prompt template. The *instruction* placeholder is replaced by the first sentence from the Evaluation Criteria column in Table 13, while *score_rubrics* is substituted with the description of each score from the same column.

For the R3 model, we employ the evaluation using the model (Anugraha et al., 2025) following the pointwise evaluation prompt template from the paper. Table 17 displays the template of the prompt. We replace these placeholders in the same way as described previously for the M-Prometheus prompt.

H Automatic Evaluation Results

This section presents the full results of the automatic evaluation using M-Prometheus (Pombal et al., 2025) and the R3 (Anugraha et al., 2025) reward model on our dataset.

Topic	Scenario Template Example 1	Scenario Template Example 2	Scenario Template Example 3
Favorite TV Shows from Childhood	Two people discuss the influence of [LANGUAGE] folklore in their favorite childhood TV shows.	Person A loved a popular [LANGUAGE] [TV_SHOWS-1], while Person B grew up watching [LANGUAGE] [TV_SHOWS-2] on TV.	Both discuss how [LANGUAGE] TV shows shaped their childhood and how modern TV differs from those days.
Favorite Musicians or Bands	Both people grew up listening to the same iconic singer, [SINGER].	Person A admires [SONG_TYPE-1] music, while Person B prefers the uniqueness of [SONG_TYPE-2].	They discuss how traditional [LANGUAGE] songs influenced their favorite [SONG_TYPE] songs nowadays.
Movie or Series Characters That Inspire You	Two people discuss how [LANGUAGE] action films' strong female leads inspired them to be more assertive in life.	Person A admires [LANGUAGE] [MOVIE_TYPE-1] movie characters, while Person B finds inspiration from modern [LANGUAGE] [MOVIE_TYPE-2] TV series.	[LANGUAGE] mythology-based movies, and how characters rooted in local legends shaped their personal values.
The Most Interesting Local Folk Tales or Myths	Both people share stories about [MYTH_CHARACTER], the [LANGUAGE] legend myth, but one believes in her protective power while the other sees her as just a legend.	Person A is fascinated by the [LANGUAGE] [MYTH_CHARACTER-1], while Person B prefers [LANGUAGE] tales of [MYTH_CHARACTER-2].	Comparing the morals behind [LANGUAGE] folk tales, focusing on [MYTH_CHARACTER-1] vs [MYTH_CHARACTER-2].
First Experience Watching a Movie in the Cinema	Two people discussing their shared excitement of watching an action movie in a small-town [LANGUAGE] cinema for the first time.	Person A was terrified by the loud sound system in a [CITY] cinema, while Person B found it thrilling and immersive.	Memorable experiences at classic [CITY] cinema chains and how they shaped their love for movies.

Table 6: Example of scenario templates used in the dialogue construction process. Templates include delexicalized placeholders (e.g., [LANGUAGE], [TV_SHOWS]) for later lexicalization.

Topic	Persona Template Example 1	Persona Template Example 2	Persona Template Example 3
Favorite TV Shows from Childhood	A person fascinated by traditional [MOVIE_TYPE] and mythological characters: [MYTH_CHARACTER]	A person who loved animated [MOVIE_TYPE] movie	A person who values [MOVIE_TYPE] TV shows
Favorite Musicians or Bands	A nostalgic [SONG_TYPE] lover who enjoys live performances	A classically trained musician who is fascinated by folk instruments: [TRADITIONAL_INSTRUMENT]	A person who enjoys discovering [MUSIC_GENRE] songs from various culture
Movie or Series Characters That Inspire You	An energetic extrovert who loves [MOVIE_TYPE]-packed movies	A thoughtful introvert who enjoys [MOVIE_TYPE]	A person who appreciates movie characters inspired by folklore and traditional values
The Most Interesting Local Folk Tales or Myths	Enthusiast of historical accuracy who loves researching the real events behind myths.	A skeptic person who enjoys listening to stories of [MYTH_CHARACTER]	A passionate storyteller who interested in myth
First Experience Watching a Movie in the Cinema	A person who likes [STATS_TYPE] movies	An adventurous moviegoer who likes [STATS_TYPE] theater	A person who likes [ENVIRONMENT_CONDITION] places

Table 7: Example of persona templates used in the dialogue construction process. Templates include delexicalized placeholders (e.g., [LANGUAGE], [TV_SHOWS]) for later lexicalization.

Delexicalized	Lexicalized
[LANGUAGE]	Thai-tha Indonesian-ind Javanese-jav Minangkabau-min Tagalog-tag Malay-mal Tamil-tam Vietnamese-vie
[TV_SHOWS]	drama-gen wayang_(puppet_show)-ind wayang_(puppet_show)-jav historical_drama-jav folklore_series-min minang_comedy-min cooking_show-tha comedy_sketch-ind mystery_thriller-gen supernatural_fable-tha variety_show-tag sitcom-tag musical_drama-mal historical_fiction-mal comedy_series-mal family_drama-mal tamil_serial-tam musical_program-tam talk_show-tam reality_show-tam cai_luong-vie tuong-vie cheo-vie ho_chi_minh_biopic-vie

Table 8: Delexicalized and Lexicalized Entities.

Personality Traits		
Active	Appreciative	Considerate
Creative	Friendly	Honest
Imaginative	Open	Patient
Witty	Ambitious	Amusing
Boyish	Businesslike	Determined

Table 9: Personality Traits.

Gender	Language	First Name	Last Name
Male	Indonesian	Andi	Setiawan
		Budi	Hidayat
		Joko	Saputra
Male	Javanese	Agus	Nugraha
		Mukhti	Wijaya
		Eko	Wicaksana
Male	Minangkabau	Zulkifli	Chaniago
		Fadli	Rasyid
		Yusuf	Putra
Male	Thai	Ananda	Chaiya
		Athit	Anuman
		Chayaphon	Bun Ma
Female	Indonesian	Nafisah	Yasmin
		Dewi	Rahayu
		Intan	Wahyuni
Female	Javanese	Maya	Whidia
		Gita	Jelita
		Kartika	Indriani
Female	Minangkabau	Nurul	Hasna
		Laila	Atiqah
		Citra	Azizah
Female	Thai	Atchara	Channarong
		Kanlaya	Kaew Buasai
		Kamala	Nunphakdi

Table 10: Sample List of Names.

[Ceremony]	[Food]
Hari_Raya-ind	ketupat-ind; rendang-min; satay-ind
Hari_Raya-jav	ketupat-ind; gudeg-jav; soto-ind
Hari_Raya-min	ketupat-ind; rendang-min; dendeng_batokok-min; ayam_pop-min
Satu_Suro-jav	gudeg-jav; nasi_liwet-jav; tongseng-jav
Loy_Krathong-tha	pad_thai-tha; green_curry-tha; mango_sticky_rice-tha
Songkran-tha	som_tam-tha; tom_yum-tha
Eid-gen	ketupat-ind; rendang-min; biryani-indic
Sham_el_Nessim-eg	ful_medames-eg; molokhia-eg
Chinese_New_Year-chi	pecking_duck-chi; xialongbao-chi; dim_sum-chi
Lantern_Festival-chi	dim_sum-chi; xialongbao-chi
Diwali-indic	biryani-indic; samosa-indic
Holi-indic	samosa-indic; biryani-indic
Pasko-tag	adobo-tag; lechon-tag
Deepavali-tam	fish_head_curry-tam; roti_prata-tam
Tet-vie	pho-vie; banh_mi-vie; goi_cuon-vie
Ramadan_markets-gen	ketupat-ind; satay-ind; rendang-min
Indonesian_Independence_Day-ind	nasi_goreng-ind; gado_gado-ind
Turun_Mandi-min	rendang-min; sate_padang-min
Kaharian_ng_Bagong_Taon-mal	nasi_lemak-mal
Banh_Chung-vie	pho-vie; banh_mi-vie
Tahun_Baru_Cina-mal	nasi_lemak-mal
Hari_raya-mal	nasi_lemak-mal; satay-mal; laksa-mal

Table 11: Mapping between FOOD and CEREMONY entites.

Prompt Template

Create a multi-turn conversation in {lang} from 2 people where the topic is: {topic_1}, and then move to the topic: {topic_2}. You must only speak in {lang}. The conversation is in a polite setting. During the conversation, the speaker calls the other with honorifics.

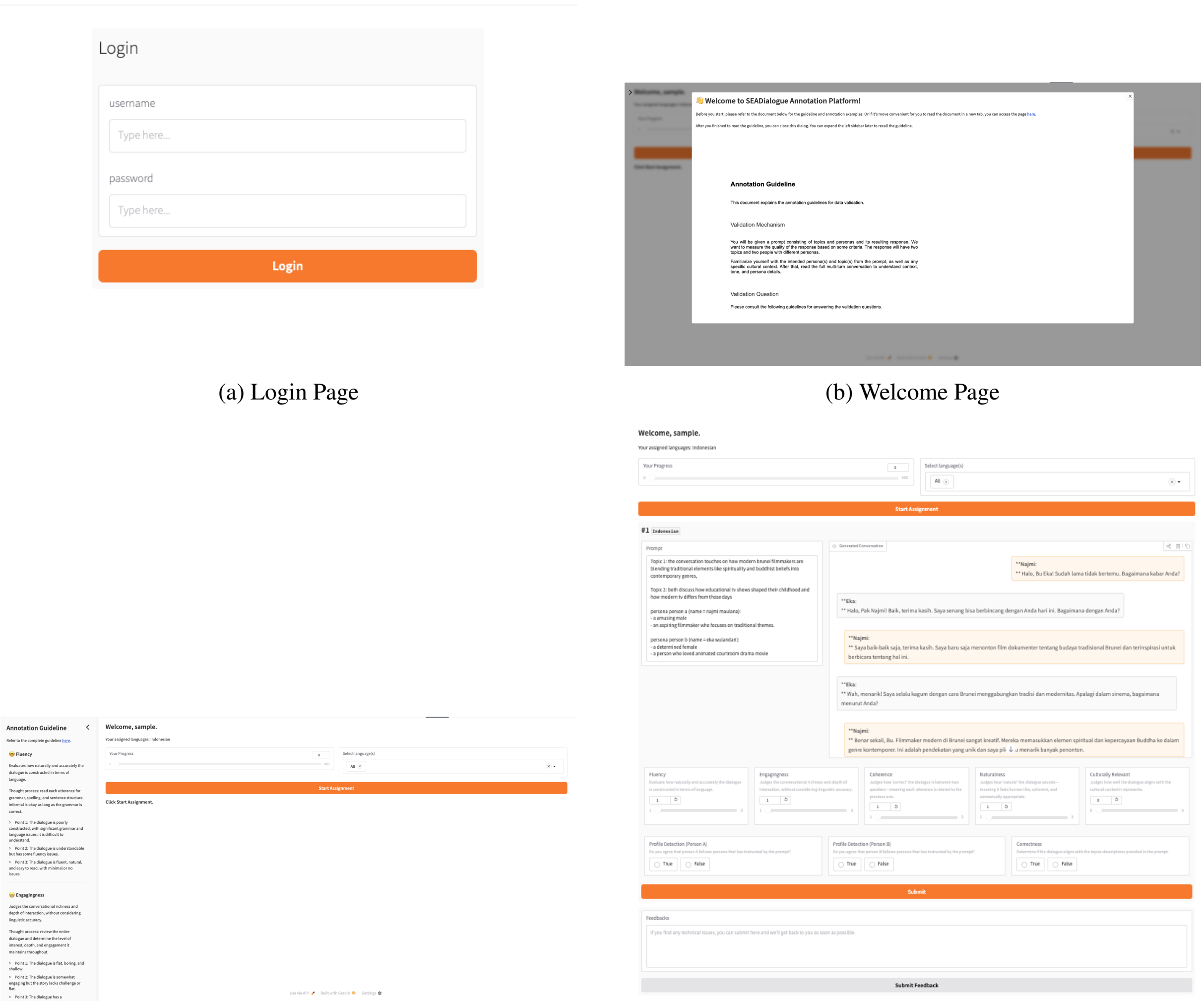
Persona Person A (name = {name_1}):
- A {personality_1} {gender_1}
- {persona_1}

Persona Person B (name = {name_2}):
- A {personality_2} {gender_2}
- {persona_2}

Limit the conversation to {num_of_turns} turns. Please be direct in generating the conversation; do not generate anything except the conversation itself. Because at least there is one topic transition in the conversation, please denote it with a special token [TRANSITION] inside the conversation. Make the transition as smooth as possible.

For every turn, please follow this format ‘name: utterance‘

Table 12: Our prompt to generate the dialogues using LLMs.



(a) Login Page

👋 Welcome to SEADialogue Annotation Platform!

Before you start, please refer to the document below for the guideline and annotation examples. Or if it's move convenient for you to read the document in a new tab, you can access the page here.

After you finished to read the guideline, you can close this dialog. You can expand the left sidebar later to recall the guideline.

Annotation Guideline

This document explains the annotation guidelines for data validation.

Validation Mechanism

You will be given a prompt consisting of topics and personas and its resulting response. We want to measure the quality of the response based on some criteria. The response will have two topics and two people with different personas.

Familiarize yourself with the intended persona(s) and topic(s) from the prompt, as well as any specific cultural context. After that, read the full multi-turn conversation to understand context, tone, and persona details.

Validation Question

Please consult the following guidelines for answering the validation questions.

(b) Welcome Page

Annotation Guideline

Refer to the complete guideline here.

🤓 Fluency

Evaluates how naturally and accurately the dialogue is constructed in terms of language.

Thought process: read each utterance for grammar, spelling, and sentence structure. Informal is okay as long as the grammar is correct.

- Point 1: The dialogue is poorly constructed, with significant grammar and language issues; it is difficult to understand.
- Point 2: The dialogue is understandable but has some fluency issues.
- Point 3: The dialogue is fluent, natural, and easy to read, with minimal or no issues.

😁 Engagingness

Judges the conversational richness and depth of interaction, without considering linguistic accuracy.

Thought process: review the entire dialogue and determine the level of interest, depth, and engagement it maintains throughout.

- Point 1: The dialogue is flat, boring, and shallow.
- Point 2: The dialogue is somewhat engaging but the story lacks challenge or flat.
- Point 3: The dialogue has a

Welcome, sample.

Your assigned languages: Indonesian

Click Start Assignment.

(c) Main Page

Welcome, sample.

Your assigned languages: Indonesian

Topic 1: the conversation touches on how modern brunei filmmakers are blending traditional elements like spirituality and buddhist beliefs into contemporary genres,

Topic 2: both discuss how educational tv shows shaped their childhood and how modern tv differs from those days

persona person a (name = najmi maulana):
- a amusing male
- an aspiring filmmaker who focuses on traditional themes.

persona person b (name = eka wulandari):
- a determined female
- a person who loved animated courtroom drama movie

**Najmi:
** Halo, Bu Eka! Sudah lama tidak bertemu. Bagaimana kabar Anda?

**Eka:
** Halo, Pak Najmi! Baik, terima kasih. Saya senang bisa berbincang dengan Anda hari ini. Bagaimana dengan Anda?

**Najmi:
** Saya baik-baik saja, terima kasih. Saya baru saja menonton film dokumenter tentang budaya tradisional Brunei dan terinspirasi untuk berbicara tentang hal ini.

**Eka:
** Wah, menarik! Saya selalu kagum dengan cara Brunei menggabungkan tradisi dan modernitas. Apalagi dalam sinema, bagaimana menurut Anda?

**Najmi:
** Benar sekali, Bu. Filmmaker modern di Brunei sangat kreatif. Mereka memasukkan elemen spiritual dan kepercayaan Buddha ke dalam genre kontemporer. Ini adalah pendekatan yang unik dan saya pik ↓ u menarik banyak penonton.

Fluency
Evaluate how naturally and accurately the dialogue is constructed in terms of language.
1 (1–3)

Engagingness
Judges the conversational richness and depth of interaction, without considering linguistic accuracy.
1 (1–3)

Coherence
Judges how 'correct' the dialogue is between two speakers—meaning each utterance is related to the previous one.
1 (1–3)

Naturalness
Judges how 'natural' the dialogue sounds—meaning it feels human-like, coherent, and contextually appropriate.
1 (1–3)

Culturally Relevant
Judges how well the dialogue aligns with the cultural context it represents.
0 (0–3)

Profile Detection (Person A)
Do you agree that person A follows persona that has instructed by the prompt?
○ True ○ False

Profile Detection (Person B)
Do you agree that person B follows persona that has instructed by the prompt?
○ True ○ False

Correctness
Determine if the dialogue aligns with the topics descriptions provided in the prompt.
○ True ○ False

If you find any technical issues, you can submit here and we'll get back to you as soon as possible.

(d) Annotation Task Assignments Display

Figure 4: Screenshots of the SEADIALOGUES Annotation Platform: (a) Login Page, (b) Welcome Page, (c) Main Page, and (d) Annotation Task Assignments Display.

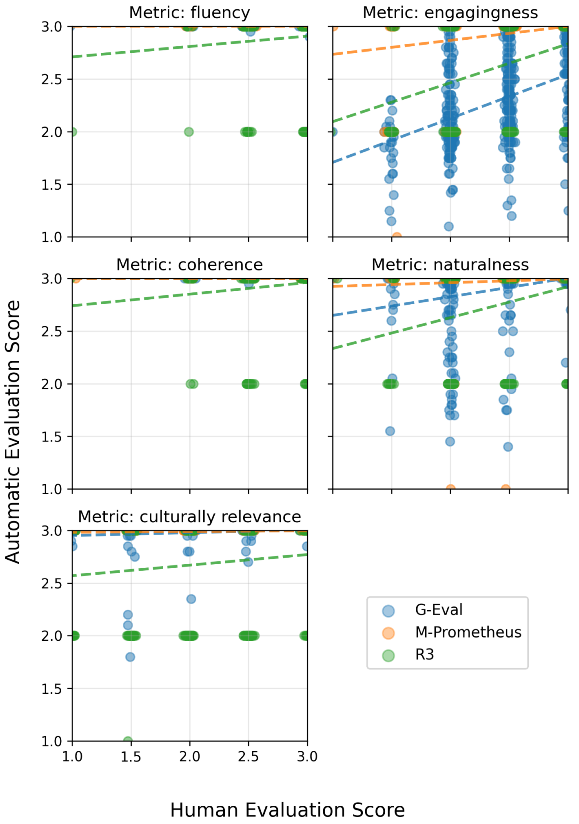


Figure 5: Visualization of the Correlation Between Automatic Evaluation and Human Annotations for All Generation Capability Metrics on Indonesia.

Table 18 displays the M-Prometheus score results for each metric, while Table 19 shows the R3 score results for every metric.

I Human-Automatic Evaluation Correlation Analysis

This section presents the supplementary figures and detailed correlation data derived from the human evaluation alongside LLM-as-judge assessments.

We extend our alignment analysis to instruction-following capabilities by binarizing human scores for the Correctness and Profile Detection metrics, which are then used as labels. The corresponding scores from the automatic evaluation serve as predictions. Based on these predictions, we compute precision, recall, F1 score, and accuracy to evaluate the reliability of the automatic evaluation. These results are reported in Tables 21, 22, 24, and 26.

Figures 5–8 further illustrate the correlation between automatic and human evaluations across the Indonesian, Minangkabau, Javanese, and Thai subsets. Tables 20, 23, and 25 report the corresponding quantitative alignment results.

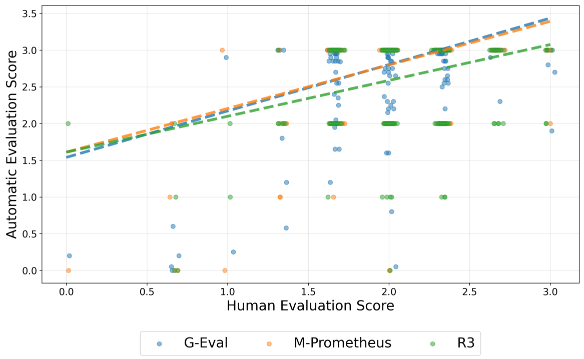


Figure 6: Visualization of the Correlation Between Automatic Evaluation and Human Annotations for Culturally Relevance Metrics on Minangkabau.

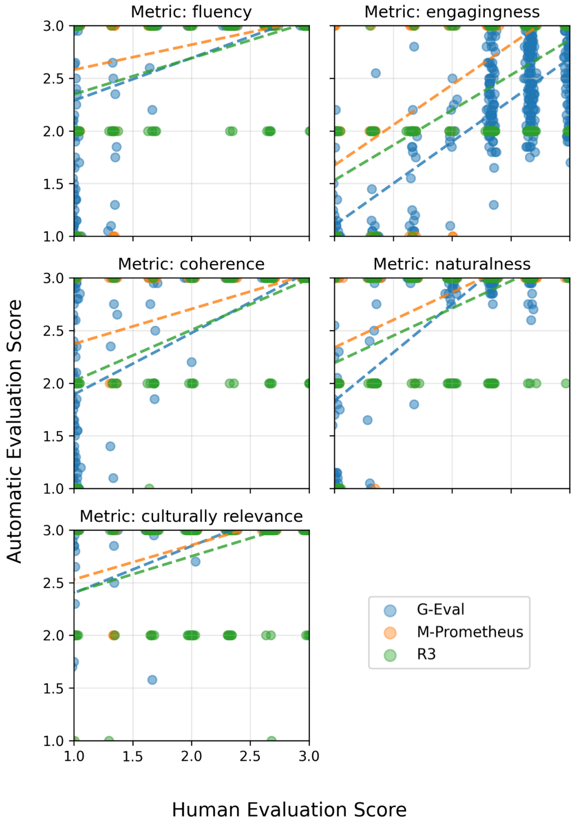


Figure 7: Visualization of the Correlation Between Automatic Evaluation and Human Annotations for All Generation Capability Metrics on Javanese.

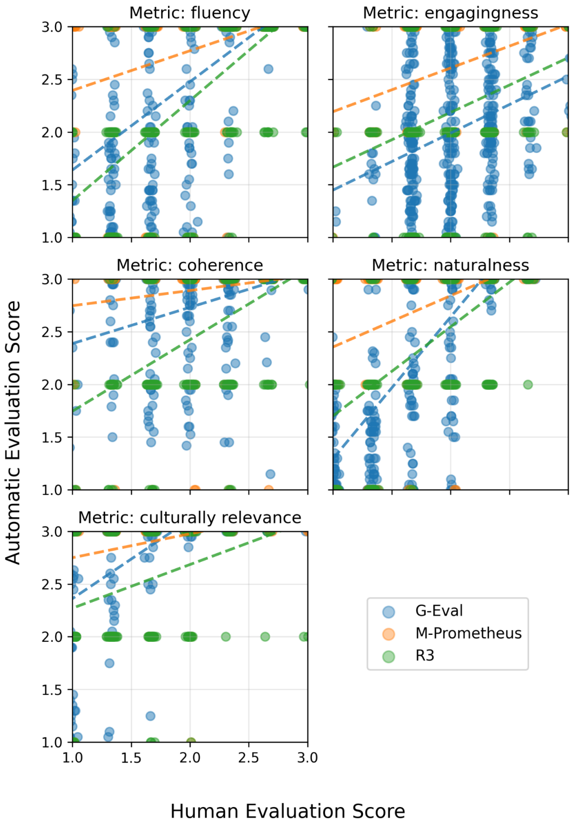


Figure 8: Visualization of the Correlation Between Automatic Evaluation and Human Annotations for All Generation Capability Metrics on Thai.

Criterion	**Score**	**Objective**	**Evaluation Criteria**
Fluency	1–3	Evaluate grammatical correctness and sentence structure.	Review each utterance for grammar, spelling, and structure. Informal language is acceptable if grammatically correct. **- 1: Poor.** The dialogue is poorly constructed, with significant grammar and language issues; it is difficult to understand. **- 2: Fair.** The dialogue is understandable but has some fluency issues. **- 3: Good**. The dialogue is fluent, natural, and easy to read, with minimal or no issues.
Engagingness	1–3	Assess the depth and interest of the conversation.	Examine the entire dialogue for richness, interaction quality, and ability to sustain interest. **- 1: Poor**. The conversation is flat, boring, and shallow. **- 2: Fair**. Somewhat engaging but the story lacks challenge or flat. **- 3: Good**. The conversation has a challenging/deep/twist topic/discussion. Living the readers to be excited.
Coherence	1–3	Ensure logical flow between utterances.	Verify that each response directly relates to the previous one. Flag any abrupt topic shifts or irrelevant responses. **- 1: Poor**. The utterances in the dialogue are completely unrelated and nonsensical. **- 2: Fair**. The dialogue is somewhat coherent, but it contains some unrelated utterances. **- 3: Good**. The dialogue is fully coherent, with no hallucinations or inconsistencies in the utterances.
Naturalness	1–3	Determine how human-like and contextually appropriate the dialogue is.	Look for natural rhythms, idiomatic expressions, and smooth transitions. Penalize mechanical or repetitive phrasing. **- 1: Poor**. The dialogue is completely unnatural and robotic, with awkward phrasing and obvious AI generation. **- 2: Fair**. The dialogue is somewhat natural but still has noticeable AI-like patterns (e.g., repetitive phrasing or awkward transitions). **- 3: Good**. The dialogue feels entirely natural, like a real human conversation, with no clear signs of AI involvement.
Cultural Relevance	0–3	Assess the accuracy of cultural references.	Identify cultural references and evaluate their correctness within the intended cultural context. **- 0:** The dialogue doesn't have cultural aspect **- 1: Poor**. The dialogue is entirely irrelevant to the intended culture, containing inaccuracies or stereotypes. **- 2: Fair**. The dialogue has some cultural relevance but includes noticeable inaccuracies or lacks depth. **- 3: Good**. The dialogue is fully culturally correct, accurately reflecting norms, knowledge, and references authentically.
Profile Detection	Y/N	Check alignment with provided persona descriptions.	Compare character behavior, tone, and language to the persona descriptions. **Y**: Traits are clearly represented. **N**: Traits are absent.
Correctness	Y/N	Verify adherence to the topic constraints.	Confirm that both specified topics from the prompt are clearly addressed. **Y**: Both topics appear. **N**: At least one topic is missing.

Table 13: Dialogue Evaluation Rubrics.

Prompt Template
You will be given one generated dialogue between two people, each with a distinct persona, discussing two predetermined topics. Your task is to rate the dialogue on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. **Evaluation Criteria:** {metric's evaluation criteria} **Evaluation Steps:** {metric's evaluation steps} **Example:** **Generated Dialogue:** {{Dialogue}} **Evaluation Form (scores ONLY):** - {metric's name}:

Table 14: Automatic Evaluation G-Eval Prompt Template.

Evaluation Steps

Fluency:

1. Read the entire dialogue thoroughly.
2. Review each utterance for grammatical correctness, spelling, and sentence structure.
3. Informal language is fine, as long as it's grammatically correct.
4. Assign a score for Fluency based on the Evaluation Criteria.

Engagingness:

1. Read the entire dialogue thoroughly.
2. Identify topic depth and richness.
3. Determine the level of interest, depth, and engagement it maintains throughout.
4. Assign a score for Engagingness based on the Evaluation Criteria.

Coherence:

1. Read the entire dialogue thoroughly.
2. Determine whether each utterance logically follows the previous one.
3. Look for any abrupt or irrelevant shifts in the conversation.
4. Assign a score for Coherence based on the Evaluation Criteria.

Naturalness:

1. Read the entire dialogue thoroughly.
2. Assess human-likeness of language. Evaluate if the wording, tone, and sentence structure feel authentic and appropriate for spoken conversation.
3. Identify AI-like artifacts. Look for robotic phrasing, overly formal or generic responses, or repeated sentence patterns.
4. Assign a score for Naturalness based on the Evaluation Criteria.

Cultural Relevance:

1. Read the entire dialogue thoroughly.
2. Determine the presence of any cultural references within the dialogue.
3. Look for cultural references in the dialogue and decide if they are accurate.
4. Assign a score for Culturally Relevance metric based on the Evaluation Criteria.

Correctness:

1. Identify the two predetermined topics.
2. Read the entire dialogue carefully.
3. Go through each utterance in the dialogue and see if it fits these topics.
4. Assign a score for Correctness based on the Evaluation Criteria.

Profile Detection:

1. Read the Persona descriptions for Person A and Person B carefully.
2. Read the entire dialogue attentively.
3. Identify whether the dialogue includes explicit or implicit elements that showcase each speaker's persona.
4. Assign score for Profile Detection based on the Evaluation Criteria for each person.

Table 15: Evaluation steps for each metric using G-EVAL.

Prompt Template

###Task Description: An instruction (might include an Input inside it), a response to evaluate, and a score rubric representing a evaluation criteria are given.
1. Write a detailed feedback that assess the quality of the response strictly based on the given score rubric, not evaluating in general.
2. After writing a feedback, write a score that is an integer between {min_score} and {max_score}. You should refer to the score rubric.
3. The output format should look as follows: "Feedback: (write a feedback for criteria) [RESULT] (an integer number between {min_score} and {max_score})"
4. Please do not generate any other opening, closing, and explanations.
###The instruction to evaluate: {instruction} The original prompt for the dialogue was: {dialogue_prompt}
###Response to evaluate: {generated_dialogue}
###Score Rubrics: {score_rubrics}
###Feedback:

Table 16: Automatic Evaluation M-Prometheus Prompt Template.

Prompt Template

Evaluate the response based on the given task, input, response, and evaluation rubric. Provide a fair and detailed assessment following the rubric.
TASK
{instruction}
INPUT
{dialogue_prompt}
RESPONSE
{generated_dialogue}
EVALUATION RUBRIC
{score_rubrics}
OUTPUT FORMAT
Return a JSON response in the following format:
{{
"explanation": "Explanation of why the response received a particular score", "score": "Score assigned to the response based on the rubric between {score_range}"
}}
EVALUATION

Table 17: Automatic Evaluation R3 Prompt Template.

Model	ind	jav	min	tha
Coherence				
Aya-8B-Expanse	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00	2.80 ± 0.53
Gemini 1.5 Flash	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00
GPT-4o mini	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00
Llama-3.1-Instruct	3.00 ± 0.00	2.42 ± 0.82	2.79 ± 0.57	2.92 ± 0.31
Culturally Relevance				
Aya-8B-Expanse	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00	2.54 ± 0.88
Gemini 1.5 Flash	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00
GPT-4o mini	3.00 ± 0.00	3.00 ± 0.00	2.99 ± 0.10	3.00 ± 0.00
Llama-3.1-Instruct	2.96 ± 0.20	2.01 ± 1.08	2.41 ± 0.83	2.66 ± 0.52
Engagingness				
Aya-8B-Expanse	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00	2.46 ± 0.63
Gemini 1.5 Flash	3.00 ± 0.00	2.99 ± 0.10	2.98 ± 0.14	2.96 ± 0.20
GPT-4o mini	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00	2.97 ± 0.17
Llama-3.1-Instruct	2.68 ± 0.49	1.68 ± 0.65	2.02 ± 0.72	2.07 ± 0.46
Fluency				
Aya-8B-Expanse	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00	2.55 ± 0.81
Gemini 1.5 Flash	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00
GPT-4o mini	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00	3.00 ± 0.00
Llama-3.1-Instruct	3.00 ± 0.00	2.17 ± 0.87	2.53 ± 0.70	2.69 ± 0.61
Naturalness				
Aya-8B-Expanse	3.00 ± 0.00	2.99 ± 0.10	2.99 ± 0.10	2.45 ± 0.78
Gemini 1.5 Flash	2.99 ± 0.10	3.00 ± 0.00	2.95 ± 0.26	2.99 ± 0.10
GPT-4o mini	2.96 ± 0.28	2.99 ± 0.10	3.00 ± 0.00	3.00 ± 0.00
Llama-3.1-Instruct	2.94 ± 0.24	1.95 ± 0.67	2.24 ± 0.67	2.34 ± 0.65
Correctness				
Aya-8B-Expanse	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.88 ± 0.33
Gemini 1.5 Flash	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.10
GPT-4o mini	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Llama-3.1-Instruct	0.98 ± 0.14	0.77 ± 0.42	0.79 ± 0.41	0.88 ± 0.33
Profile Detection				
Aya-8B-Expanse	0.98 ± 0.14	0.99 ± 0.07	0.99 ± 0.07	0.79 ± 0.41
Gemini 1.5 Flash	0.99 ± 0.10	1.00 ± 0.00	1.00 ± 0.00	0.98 ± 0.12
GPT-4o mini	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.07	0.98 ± 0.14
Llama-3.1-Instruct	0.98 ± 0.14	0.83 ± 0.38	0.95 ± 0.22	0.92 ± 0.27

Table 18: M-Prometheus results on model's conversational generation and instruction following capabilities.

Model	ind	jav	min	tha
Coherence				
Aya-8B-Expanse	2.94 ± 0.24	2.89 ± 0.31	2.94 ± 0.24	1.85 ± 0.36
Gemini 1.5 Flash	3.00 ± 0.00	2.97 ± 0.17	2.98 ± 0.14	2.96 ± 0.20
GPT-4o mini	2.99 ± 0.10	2.99 ± 0.10	2.98 ± 0.14	2.96 ± 0.20
Llama-3.1-Instruct	2.83 ± 0.38	2.05 ± 0.54	2.53 ± 0.58	2.68 ± 0.47
Culturally Relevance				
Aya-8B-Expanse	2.67 ± 0.49	2.74 ± 0.46	2.65 ± 0.59	1.71 ± 0.50
Gemini 1.5 Flash	2.83 ± 0.38	2.96 ± 0.20	2.92 ± 0.27	2.79 ± 0.41
GPT-4o mini	2.71 ± 0.46	2.93 ± 0.26	2.69 ± 0.54	2.73 ± 0.47
Llama-3.1-Instruct	2.51 ± 0.52	1.94 ± 0.55	2.26 ± 0.61	2.22 ± 0.48
Engagingness				
Aya-8B-Expanse	2.77 ± 0.42	2.71 ± 0.46	2.76 ± 0.43	1.48 ± 0.50
Gemini 1.5 Flash	2.76 ± 0.43	2.61 ± 0.49	2.64 ± 0.48	2.69 ± 0.46
GPT-4o mini	2.75 ± 0.44	2.67 ± 0.47	2.54 ± 0.50	2.60 ± 0.49
Llama-3.1-Instruct	2.17 ± 0.38	1.60 ± 0.53	1.93 ± 0.43	2.01 ± 0.27
Fluency				
Aya-8B-Expanse	2.92 ± 0.27	2.85 ± 0.36	2.80 ± 0.43	1.40 ± 0.49
Gemini 1.5 Flash	2.98 ± 0.14	2.96 ± 0.20	2.93 ± 0.26	2.92 ± 0.27
GPT-4o mini	2.92 ± 0.27	2.95 ± 0.22	2.83 ± 0.43	2.97 ± 0.17
Llama-3.1-Instruct	2.76 ± 0.43	1.86 ± 0.53	2.18 ± 0.54	2.30 ± 0.46
Naturalness				
Aya-8B-Expanse	2.81 ± 0.39	2.85 ± 0.36	2.72 ± 0.45	1.58 ± 0.50
Gemini 1.5 Flash	2.97 ± 0.17	2.83 ± 0.38	2.87 ± 0.34	2.73 ± 0.45
GPT-4o mini	2.89 ± 0.31	2.81 ± 0.39	2.80 ± 0.40	2.78 ± 0.42
Llama-3.1-Instruct	2.29 ± 0.46	1.85 ± 0.44	2.04 ± 0.37	2.08 ± 0.27
Correctness				
Aya-8B-Expanse	0.90 ± 0.30	0.92 ± 0.27	0.95 ± 0.22	0.13 ± 0.34
Gemini 1.5 Flash	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
GPT-4o mini	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.10	1.00 ± 0.00
Llama-3.1-Instruct	0.89 ± 0.31	0.55 ± 0.50	0.75 ± 0.44	0.73 ± 0.45
Profile Detection				
Aya-8B-Expanse	0.89 ± 0.32	0.88 ± 0.33	0.92 ± 0.27	0.31 ± 0.46
Gemini 1.5 Flash	0.85 ± 0.36	0.88 ± 0.33	0.86 ± 0.35	0.80 ± 0.40
GPT-4o mini	0.84 ± 0.37	0.88 ± 0.33	0.87 ± 0.34	0.79 ± 0.41
Llama-3.1-Instruct	0.76 ± 0.43	0.41 ± 0.49	0.66 ± 0.47	0.66 ± 0.47

Table 19: R3 results on model’s conversational generation and instruction following capabilities.

	G-Eval	**M-Prometheus**	**R3**
Pearson			
Coherence	0.0555	NaN	**0.1336**
Culturally Relevance	**0.1704**	0.0489	0.1508
Engagingness	**0.4282**	0.2093	0.3353
Fluency	0.0123	NaN	**0.0795**
Naturalness	0.2429	0.0733	**0.2658**
Spearman			
Coherence	0.0684	NaN	**0.1440**
Culturally Relevance	**0.1846**	0.0587	0.1587
Engagingness	**0.4333**	0.1841	0.3249
Fluency	**0.0460**	NaN	0.0397
Naturalness	**0.3188**	0.0835	0.2690
KendallTau			
Coherence	0.0673	NaN	**0.1417**
Culturally Relevance	**0.1633**	0.0527	0.1424
Engagingness	**0.3440**	0.1711	0.3020
Fluency	**0.0456**	NaN	0.0394
Naturalness	**0.2835**	0.0783	0.2529

Table 20: Automatic Evaluations and Human Annotations Correlations on Indonesian. M-Prometheus assigns the same values to both Coherence and Fluency, making correlation calculation not possible (resulting in NaN).

LLM Judge	**P**	**R**	**F1**	**Accuracy**
G-Eval				
Correctness	0.91	**0.99**	0.94	**0.91**
Profile Detection	0.41	**0.99**	0.58	0.42
M-Prometheus				
Correctness	0.91	0.99	**0.95**	0.91
Profile Detection	0.41	0.98	0.57	0.41
R3				
Correctness	**0.92**	0.96	0.94	0.90
Profile Detection	**0.46**	0.94	**0.61**	**0.52**

Table 21: Comparison of Automatic Evaluation Predictions to Human Annotations for Instruction Following Metrics on Indonesian.

LLM Judge	**P**	**R**	**F1**	**Accuracy**
G-Eval				
Correctness	0.985	**0.987**	**0.986**	**0.972**
Profile Detection	0.977	0.966	0.971	0.945
M-Prometheus				
Correctness	**0.986**	0.956	0.971	0.945
Profile Detection	0.973	**0.990**	**0.982**	**0.965**
R3				
Correctness	0.983	0.928	0.955	0.915
Profile Detection	**0.981**	0.837	0.903	0.828

Table 22: Comparison of Automatic Evaluation Predictions to Human Annotations for Instruction Following Metrics on Minangkabau.

	G-Eval	**M-Prometheus**	**R3**
Pearson			
Coherence	**0.7899**	0.5496	0.7794
Culturally Relevance	**0.6228**	0.5557	0.5956
Engagingness	**0.8338**	0.8377	0.6935
Fluency	0.5007	0.3743	**0.5105**
Naturalness	**0.6811**	0.5668	0.5424
Spearman			
Coherence	**0.7846**	0.4864	0.7040
Culturally Relevance	**0.6893**	0.5551	0.5956
Engagingness	0.6684	**0.7180**	0.5878
Fluency	0.5022	0.3794	**0.5036**
Naturalness	0.5979	**0.6319**	0.5731
KendallTau			
Coherence	**0.7022**	0.4505	0.6518
Culturally Relevance	**0.5775**	0.4789	0.4868
Engagingness	0.5372	**0.6364**	0.5235
Fluency	0.4426	0.3415	**0.4535**
Naturalness	0.4867	**0.5483**	0.4983

Table 23: Automatic Evaluations and Human Annotations Correlations on Javanese.

LLM Judge	**P**	**R**	**F1**	**Accuracy**
G-Eval				
Correctness	0.846	**0.996**	0.916	0.853
Profile Detection	0.772	**1.000**	**0.871**	0.795
M-Prometheus				
Correctness	0.846	0.991	0.912	0.847
Profile Detection	0.724	0.998	0.839	0.734
R3				
Correctness	**0.893**	0.966	**0.928**	**0.880**
Profile Detection	**0.822**	0.903	0.861	**0.797**

Table 24: Comparison of Automatic Evaluation Predictions to Human Annotations for Instruction Following Metrics on Javanese.

	G-Eval	M-Prometheus	R3
Pearson			
Coherence	0.4191	0.2116	**0.5655**
Culturally Relevance	**0.4994**	0.2293	0.3596
Engagingness	**0.3353**	0.2694	0.2842
Fluency	0.6591	0.3751	**0.6986**
Naturalness	**0.7408**	0.3751	0.5987
Spearman			
Coherence	0.5079	0.2092	**0.5482**
Culturally Relevance	**0.5146**	0.2168	0.3798
Engagingness	**0.3323**	0.2657	0.2805
Fluency	0.6995	0.3598	**0.7001**
Naturalness	**0.7251**	0.3851	0.5952
KendallTau			
Coherence	0.4279	0.1859	**0.4877**
Culturally Relevance	**0.4214**	0.1889	0.3267
Engagingness	**0.2519**	0.2395	0.2485
Fluency	0.5778	0.3154	**0.6103**
Naturalness	**0.5941**	0.3388	0.5221

Table 25: Automatic Evaluations and Human Annotations Correlations on Thai.

LLM Judge	P	R	F1	Accuracy
G-Eval				
Correctness	0.8630	0.9642	**0.9108**	0.8550
Profile Detection	0.5417	0.9487	0.6897	0.5838
M-Prometheus				
Correctness	0.7973	**0.9739**	0.8768	0.7900
Profile Detection	0.5109	**0.9615**	0.6673	0.5325
R3				
Correctness	**0.9406**	0.8762	0.9073	**0.8625**
Profile Detection	**0.6169**	0.8051	**0.6986**	**0.6608**

Table 26: Comparison of Automatic Evaluation Predictions to Human Annotations for Instruction Following Metrics on Thai.