PINGPONG: A Natural Benchmark for Multi-Turn Code-Switching Dialogues

Anonymous ACL submission

Abstract

Code-switching is a widespread practice among the world's multilingual majority, yet few benchmarks accurately reflect its complexity in everyday communication. We present PINGPONG, a benchmark for natural multi-party code-switching dialogues covering five language-combination variations, some of which are trilingual. Our dataset consists of human-authored conversations among 2 to 4 participants covering authentic, multi-threaded structures where replies frequently reference much earlier points in the dialogue. We demonstrate that our data is significantly more natural and structurally diverse than machine-generated alternatives, offering greater variation in message length, speaker dominance, and reply distance. Based on these dialogues, we define three downstream tasks: QUESTION ANSWERING, DIALOGUE SUMMARIZATION, and TOPIC CLASSIFICATION. Evaluations of several state-of-the-art language models on PINGPONG reveal that performance remains limited on code-switched inputs, underscoring the urgent need for more robust NLP systems capable of addressing the intricacies of real-world multilingual discourse.

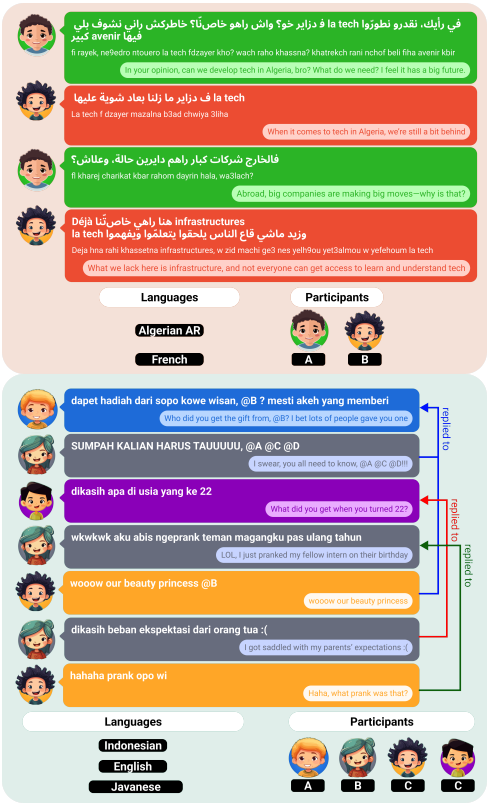


Figure 1: Illustration of dialogue samples from PINGPONG. The bottom panel highlights how each turn corresponds to previous turns, as labeled in our dataset.

1 Introduction

Code-switching, the practice of alternating between two or more languages within a single conversation, is a pervasive linguistic phenomenon in contemporary multilingual societies (Poplack, 2013; Myers-Scotton, 1997; Bullock and Toribio, 2009; Auer, 2013). With the global number of multilingual speakers surpassing that of monolinguals, the frequency of code-switching in both informal and formal communication continues to rise (Tucker, 1999; Grosjean, 2010; Wei, 2018; Winata et al., 2021). Despite recent progress in multilingual language models (Xue et al., 2020; Aryabumi et al., 2024; Team et al., 2025; Yang et al., 2025), their ability to handle code-switched dialogue remains underexplored and insufficiently evaluated (Li and Fung, 2012; Winata et al., 2018; Adilazuarda et al., 2022). Although several code-switching benchmarks have been introduced in the past, these resources are becoming outdated, and therefore, increasingly misaligned with the current era of LLMs and are less reflective of the everyday realities of code-switching (Aguilar et al., 2020b; Khanuja et al., 2020b).

To address this gap, we present PINGPONG, a novel benchmark for code-switching in dialogues. Our benchmark consists of conversations in which

Dataset	Open-Source	CS Type	#Comp. Lang.	#Regions	#Gen. Task	Conv.
SEAME (Lyu et al., 2010)	✓	Bilingual	0	1	0	✓
GLUECoS (Khanuja et al., 2020a)	✓	Bilingual	0	2	1	✗
CommonDost (Parekh et al., 2020)	✓	Bilingual	0	2	1	✓
MalayalamMixSentiment (Chakravarthi et al., 2020a)	✓	Bilingual	1	1	0	✗
TamilMixSentiment (Chakravarthi et al., 2020b)	✗	Bilingual	0	1	0	✗
hinglishNorm (Makhija et al., 2020)	✓	Bilingual	0	1	1	✓
LinCE (Aguilar et al., 2020a)	✓	Bilingual	2	3	1	✗
CSCS (Balabel et al., 2020)	✓	Bilingual	1	1	0	✗
CanVEC (Nguyen and Bryant, 2020)	✓	Bilingual	1	1	1	✗
ArzEn (Hamed et al., 2020)	✗	Bilingual	1	1	1	✗
PHINC (Srivastava and Singh, 2020)	✓	Bilingual	0	1	1	✗
MIPE (Garg et al., 2021)	✗	Bilingual	0	1	1	✗
HinGE (Srivastava and Singh, 2021)	✓	Bilingual	0	1	1	✗
TCS (Tarunesh et al., 2021)	✗	Bilingual	0	1	2	✓
GupShup (Mehnaz et al., 2021)	✓	Bilingual	0	1	1	✓
DOSA (Ravikiran and Annamalai, 2021)	✓	Bilingual	1	1	0	✗
TweetTaglish (Herrera et al., 2022)	✓	Bilingual	1	1	0	✗
BaSCo (Aguirre et al., 2022)	✓	Bilingual	1	1	0	✓
MHE (Rani et al., 2022)	✗	Trilingual	1	1	0	✗
ASCEND (Lovenia et al., 2022)	✓	Bilingual	0	1	1	✓
L3Cube-HingCorpus (Nayak and Joshi, 2022)	✓	Bilingual	0	1	1	✗
CroCoSum (Zhang and Eickhoff, 2024)	✓	Bilingual	0	1	1	✗
CS-PREF (Kuwanto et al., 2024)	✓	Bilingual	1	2	1	✗
CS-Sum (Suresh et al., 2025)	✓	Bilingual	1	2	1	✗
CodeMixQA (Winata et al., 2026)	✓	Bilingual	16	3	1	✗
PINGPONG	✓	Bilingual Trilingual	**4**	**3**	**2**	✓

Table 1: Comparison of code-switching benchmark datasets. **Open-Source** indicates public availability of the dataset. **CS Type** denotes the code-switching setting. **# Comp. Lang.** counts under-studied languages beyond high-resource ones (e.g., English, Hindi, Chinese). **# Regions** reports geographic coverage (e.g., South Asian-English, European-English). **# Gen. Tasks** indicates the number of generative tasks. **Conv.** denotes whether the dataset is conversational.

fluent multilingual speakers naturally code-switch between languages while discussing predefined topics, resulting in multiple language-combination variations through multi-party conversational interactions. PINGPONG comprises three downstream tasks, including QUESTION ANSWERING, DIALOGUE SUMMARIZATION, and MULTI-LABEL TOPIC CLASSIFICATION. In addition, the benchmark provides broad language coverage, spanning both high-resource and low-resource language combinations, with certain language combination further distinguished by differences in writing scripts.

In summary, our contributions are as follow:

- We introduce PINGPONG [1], a novel benchmark for evaluating natural multi-party code-switching in conversational settings, designed to capture the fluid dynamics of multilingual communication.
- PINGPONG consists of authentic interactions among 2 to 4 speakers discussing everyday topics, with conversations ranging from brief exchanges of 17 turns to extended dialogues of up to 189 turns. Reflecting real conversational behavior, speakers may produce long-distance replies by responding to earlier messages while alternating turns.
- PINGPONG covers five language combinations spanning high-resource, low-resource, and diverse-script languages, enabling evaluation across a broad range of multilingual scenarios.
- Our evaluation framework spans three downstream tasks—QUESTION ANSWERING, DIALOGUE SUMMARIZATION, and TOPIC CLASSIFICATION—providing a holistic assessment of language understanding and generation under code-switching conditions.
- We conduct extensive experiments with both

[1] Data will be released publicly

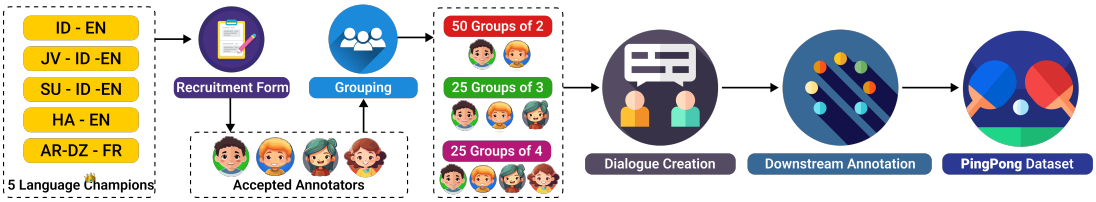


Figure 2: Overview of PINGPONG dataset construction. We recruit annotators who speak five language combinations and group them into small teams of 2 to 4 participants. Each group engages in multi-party dialogue creation, after which downstream annotations are performed. The final output is the curated PINGPONG dataset, focusing on natural code-switched dialogues for multiple tasks.

open-source and proprietary LLMs, revealing their limitations and highlighting persistent challenges in modeling code-switched, multi-party conversations.

2 Related Work

Research on code-switching has been studied extensively by linguists for decades (Sitaram et al., 2019). In NLP, code-switching has been explored across a wide range of tasks (Winata et al., 2023), including language identification (Aguilar and Solorio, 2020; Burchell et al., 2024; Ojo et al., 2025; Das et al., 2023), named entity recognition (Winata et al., 2019a; Aguilar et al., 2018; Jain et al., 2018), part-of-speech tagging (Soto and Hirschberg, 2018; Çetinoğlu and Çöltekin, 2016; Ball and Garrette, 2018), and sentiment analysis (Zhang et al., 2021; Shynkarov et al., 2025). In parallel, dataset development and shared evaluations have been actively promoted through initiatives such as CALCS (Heredia et al., 2025; Babatunde et al., 2025) and the FIRE shared tasks (Kulkarni et al., 2024; Adeyemi et al., 2023; Mandl et al., 2024). More recently, this line of work has begun to consolidate around shared benchmarks that enable systematic comparison across models, language combinations, and tasks.

However, despite the growing body of work, several limitations remain across existing code-switching datasets (Table 1). From a sustainability perspective, a number of datasets are not publicly released (Chakravarthi et al., 2020b; Hamed et al., 2020; Garg et al., 2021). In terms of linguistic diversity, most benchmarks focus exclusively on bilingual code-switching, with only a single dataset explicitly addressing trilingual scenarios (Rani et al., 2022). Moreover, both the language and regional coverage remain narrow, as code-switching phenomena in many languages are still under-explored in NLP (Chakravarthi et al., 2020a; Aguilar et al., 2020a; Balabel et al., 2020; Nguyen and Bryant, 2020; Hamed et al., 2020; Ravikiran and Annamalai, 2021; Herrera et al., 2022; Aguirre et al., 2022; Rani et al., 2022; Kuwanto et al., 2024; Suresh et al., 2025; Winata et al., 2026).

In addition, most existing datasets fall short of capturing the conversational nature of real-world code-switching, which often emerges in spontaneous and interactive settings (Lyu et al., 2010; Parekh et al., 2020; Makhija et al., 2020; Tarunesh et al., 2021; Mehnaz et al., 2021; Lovenia et al., 2022). Evaluation tasks are also frequently limited to traditional benchmarks that no longer reflect contemporary NLP challenges (Khanuja et al., 2020a; Nayak and Joshi, 2022). To address these gaps, we introduce PINGPONG, a large-scale open-source evaluation suite that expands language and regional coverage, incorporates diverse and realistic conversational code-switching scenarios, and emphasizes generative and semantic tasks.

3 PINGPONG Dataset

Our dataset is constructed through manual crowdsourcing (Figure 2), with dialogues collected and downstream tasks curated by native speakers of each language combination. In PINGPONG, the covered language combinations include five combinations: Indonesian–English (ID-EN), Sundanese–Indonesian–English (SU-ID-EN), Javanese–Indonesian–English (JV-ID-EN), Hausa–English (HA-EN), and Algerian Arabic–Standard Arabic–French (AR-DZ-FR). This design ensures both linguistic fidelity and overall resource quality. Consequently, our benchmark provides a more reliable evaluation setting compared to datasets generated automatically by LLMs.

3.1 Annotator Hiring

We initiate the annotator recruitment process by first identifying the most prevalent language combinations exhibiting code-switching phenomena. For each language combination, we recruit a language

champion—a native speaker responsible for managing data collection for that combination. The language champions then coordinate the recruitment of annotators to contribute to dataset construction.

3.1.1 Recruitment Form

Each language champion prepares a recruitment form. This form serves two primary purposes: (i) collecting demographic information from prospective annotators, and (ii) filtering respondents into the final pool of selected annotators. The form is structured into several sections, including an introduction, self-assessment, and language assessment. Further details regarding the recruitment form and collected annotator demographics are provided in Appendix A and B, respectively.

3.1.2 Grouping

For each selected annotator, we organize dialogue collection groups using the Discord[2] platform. Given the limited number of recruited annotators (detailed guidelines are provided in Appendix C), each individual participates in multiple groups. In total, we establish 100 dialogue collection groups for each language combination, comprising 50 groups with two speakers, 25 groups with three speakers, and 25 groups with four speakers. To support this process, we notify group members at three key points: five minutes before the session, at the start of the session, and at its conclusion. Annotators who are not directly involved in dialogue collection are subsequently assigned to annotate the downstream task datasets.

3.2 Dataset Creation

3.2.1 Conversational Dialogue Collection

We conduct a separate session for each assigned annotation group. In each session, participants engage in a 15-minute conversation on a pre-assigned topic. To capture natural code-switching behavior, participants are encouraged to mix languages within the designated language combinations whenever it feels natural to do so. Code-switching is allowed at multiple levels, including individual words, sentences, paragraphs, or even within a single utterance. To preserve anonymity, participants are asked not to refer to one another by name, whether in full, abbreviated, or nickname form. Instead, they are instructed to use Discord's mention feature whenever direct reference to another participant is necessary. This procedure ensures that all participant identities remain anonymized upon data export. Additionally, participants are allowed to use Discord's reply feature to respond directly to specific utterances, thereby maintaining a coherent dialogue structure.

3.2.2 Question Answering

Once the dialogues are collected, annotators write up to 10 question-answering items for each of the corresponding dialogue, these consisted of up to 5 answerable questions and 5 unanswerable questions. The answerable questions were designed to emphasize reasoning, such that answers could not be obtained directly from the dialogue. Instead, they required the use of external knowledge (primarily knowledge specific to particular cultural regions or discussed topics, where people who lived there or familiar with the topic would be familiar) and structured reasoning to reach a correct answer. The unanswerable questions, on the other hand, were constructed following five categories—Negation, Antonym, Entity–Swap, Mutual-Exclusion, and Impossible Condition—as defined in SQuAD 2.0 or SQuADRUn (Rajpurkar et al., 2018). All QA items were formulated as multiple-choice questions with five options, where option *E* explicitly denoted "*No correct answer*". Furthermore, each question was written in the designated first language (L1) of the corresponding language combination. For example, for the Javanese–Indonesian–English dataset, Indonesian was used as the language for the QA construction.

3.2.3 Dialogue Summarization

For each dialogue, we collect up to 3 distinct summaries, each from different annotators. Each summary is 3-5 sentences long and written in the designated L1 of the corresponding language combination. Annotators were provided with the initial topic of the dialogue and instructed to only summarize relevant information, excluding any off-topic segments. We encouraged annotators to follow the four dimensions of summarization quality defined in (Kryscinski et al., 2019): **Coherence**—The logical flow and collective quality of all sentences in the summary, **Fluency**—The grammatical quality of each individual sentence, **Relevance**—The inclusion of only important, on-topic information, and **Consistency**—The degree to which all facts in the summary are supported by the source document. Table 2 presents the pairwise inter-annotator agreement scores for the written summaries, measured

[2] https://discord.com/.

Lang	Annotator 1–2	Annotator 1–3	Annotator 2–3
ID-EN	0.884	0.891	0.882
JV-ID-EN	0.891	0.888	0.884
SU-ID-EN	0.885	0.895	0.886
ARQ-FRA	0.858	–	–
HA-EN	0.883	0.887	0.880

Table 2: Pairwise BERTScore F1 between annotator-written summaries as an inter-annotator agreement score across different language combinations (Lang).

Lang	#Dialogue	#Sci/Tech	#Ent	#Soc/Cul	#Edu	#Daily
ID-EN	100	13	21	32	15	19
JV-ID-EN	100	4	17	22	17	40
SU-ID-EN	100	7	18	48	9	18
HA-EN	100	4	15	25	5	51
AR-DZ-FR	100	14	21	21	24	20

Table 3: Topic distribution of the different dialogues per language combination (Lang). We show the total number of dialogues (#Dialogue) and per topic—science and technology (#Sci/Tech), entertainment (#Ent), society and culture (#Soc/Cul), education (#Edu), and daily life (#Daily).

using BERTScore F1.

3.2.4 Topic Classification

Since the dialogue were collected based on a predefined initial topic, therefore naturally we can convert our dataset into a topic classification one as well. Specifically, we map the initial topics into classes as follow: **Science/Technology** (e.g., scientific breakthroughs, research, new technology, etc.), **Entertainment** (Sports, Music, Tourism, and any other form of entertainments), **Social/Culture** (e.g., languages, work culture, customs, traditions, etc.) **Education** (e.g., curriculum, schools, colleges, etc.), and **Daily Life** (e.g., personal experiences, love stories, daily habits). Table 3 shows the distribution of the topics in the dataset.

3.3 Data Statistics

The statistics of dialogues in PINGPONG are summarized in Table 4. Our dataset comprises long, spontaneous conversations between 2 to 4 participants, providing a representative sample of authentic human interactions. As such, it is uniquely suited for benchmarking models on long-context, natural code-switching behavior. To quantify the linguistic complexity of these dialogues, we report two standard metrics: **Code-Mixing Index (CMI)** and **Switch Point Fraction (SPF)**.

Code-Mixing Index (CMI). CMI (Gambäck and Das, 2016) measures the intensity of code-switching by assessing the distribution of languages within an utterance. While the original formulation suggests treating Named Entities separately to avoid an artificial inflation of the index, robust NER tools are often unavailable for the under-resourced languages in our study. Consequently, following Winata et al. (2019b), we simplify the calculation by not distinguishing NEs.

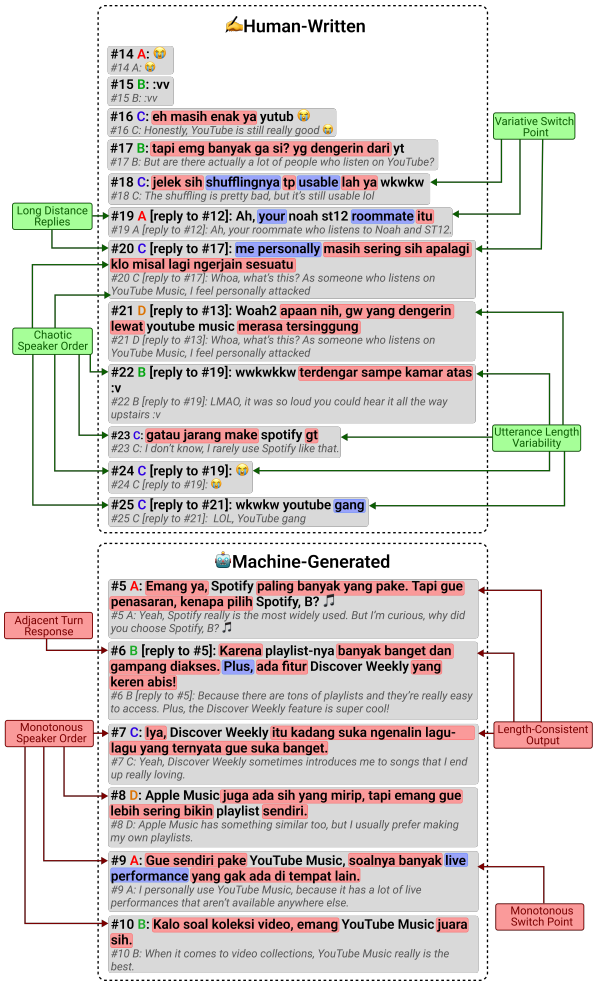


Figure 3: Comparison between human-written and machine-generated conversation texts. Text with a ■ color denotes the original conversation, while text with a ■ color represents the English translation. Segments highlighted in ■ indicate Indonesian words, whereas segments highlighted in ■ indicate English words.

Switch Point Fraction (SPF). While CMI captures the volume of mixing, SPF (Pratapa et al., 2018) focuses on the frequency of transitions between languages. It is calculated as the ratio of the number of language-switch points to the total number of possible switching positions.

3.4 Natural Conversational Pattern

Our dataset reflects the organic characteristics of human text-based communication, particularly in

Lang.	#Dialogue	Avg. Turn	Avg. Words	Avg. Tokens	Avg. CMI	Avg. SPF
ID-EN	100	81.93	448.96	782.15	0.472	0.300
JV-ID-EN	100	83.07	410.14	754.04	0.467	0.318
SU-ID-EN	100	59.88	494.03	957.20	0.757	0.467
HA-EN	100	69.14	421.32	649.84	0.352	0.190
AR-DZ-FR	100	98.60	462.74	1,120.81	0.568	0.253

Table 4: **Dialogue data statistics.** For each language combination (Lang.), we report the average number of turns (Avg. Turn), average number of words (Avg. Words), average number of tokens (Avg. Tokens), average Code-Mixing Index (Avg. CMI), and average Switch-Point Fraction (Avg SPF) per dialogue.

multi-party settings. We observe significant participation imbalance, where certain speakers dominate the conversation while others remain less active. Furthermore, human dialogue often exhibits multi-threaded structures; speakers frequently reply to refer and respond to messages from several turns prior. Utterance length is also highly variable, ranging from detailed explanations to brief one- or two-word reactions. Notably, human participants sometimes send multiple consecutive messages before a turn change occurs.

Metric	Human-written	Machine-generated
2 speakers (50 dialogue)		
Avg length variance (tokens)	66.540	35.140
Total replies	658.2	2.6
Avg degree of reply distance	2.729	0.048
Avg imbalance ratio of speaker turns	1.366	1.019
Avg CMI	0.525	0.541
Avg SPF	0.306	0.313
Human preference	2.727	2.178
3 speakers (25 dialogue)		
Avg length variance (tokens)	59.093	30.230
Total replies	630.4	19.8
Avg degree of reply distance	3.698	0.738
Avg imbalance ratio of speaker turns	2.961	1.056
Avg CMI	0.521	0.486
Avg SPF	0.305	0.284
Human preference	2.720	2.141
4 speakers (25 dialogue)		
Avg length variance (tokens)	81.195	29.191
Total replies	869.6	45.0
Avg degree of reply distance	4.159	1.085
Avg imbalance ratio of speaker turns	3.256	1.118
Avg CMI	0.520	0.483
Avg SPF	0.306	0.284
Human preference	2.676	2.078

Table 5: Comparison of conversational statistics between human-written and machine-generated dialogues, averaged across all language combinations. Human preference scores are measured using a 1–3 Likert scale.

In contrast, machine-generated dialogues produced by GPT-4o based on the same topics appear significantly more monotonous. These conversations tend to follow a rigid turn-taking structure with a uniform distribution among speakers and consistent utterance lengths. Synthetic dialogues rarely feature consecutive messages from a single speaker, and replies are almost always linear and immediate. A comparison of these patterns is illustrated in Figure 3. To quantify these observations, we provide statistical analysis in Table 5, which shows that across most languages, human annotators consistently rate organic conversations as more natural than their machine-generated counterparts. We use Free-marginal multirater kappa (Randolph, 2010) to calculate the inter-annotator agreement on the annotations, and got 0.172 across all data points. This slight agreement highlights the inherently subjective nature of evaluating conversational naturalness. Detailed statistics for each language combination can be found in Appendix H.3.

4 Experimental Setup

We evaluate a diverse set of models across the three tasks. From the perspective of language coverage, these models can be categorized into English-centric models, multilingual models, and region-specific models tailored to particular language combinations. We also compare base models with reasoning-oriented and instruction-tuned variants.

Regarding linguistic coverage, our study leverages models such as SAILOR2 (Dou et al., 2025), AYA23 (Aryabumi et al., 2024), SAHABAT AI,[3] QWEN2.5 (Qwen et al., 2025), GEMMA2 (Team et al., 2024), GEMMA3 (Team et al., 2025), ALLAM (Bari et al., 2025), and SILMA (silma-ai, 2024). Furthermore, to assess reasoning capabilities, we include QWEN3 (Yang et al., 2025) models with 4B and 8B parameters. Detailed hyperparameter configurations for each model are provided in Appendix 8. We do single-run for our experiment due to the cost. Moreover, our novelty is in the dataset and we do not claim definite ranking on the models, rather we show general performance across diverse set of models and setting in our dataset.

4.1 Reasoning Behavior

Some of our models, namely the Qwen3 series, have reasoning capabilities. Therefore, we inves-

[3] https://huggingface.co/Sahabat-AI.

	QUESTION ANSWERING (ACCURACY %)					DIALOGUE SUMMARIZATION (BERTSCORE-F)					TOPIC CLASSIFICATION (ACCURACY %)				
Model	ID-EN	JV-ID-EN	SU-ID-EN	HA-EN	AR-DZ-FR	ID-EN	JV-ID-EN	SU-ID-EN	HA-EN	AR-DZ-FR	ID-EN	JV-ID-EN	SU-ID-EN	HA-EN	AR-DZ-FR
Global															
Qwen2.5-3B-Instruct	5.86	18.98	76.16	43.93	22.22	0.894	0.893	0.901	0.777	0.784	52.53	38.38	52.53	30.30	46.46
Qwen2.5-7B-Instruct	4.64	18.98	88.68	43.93	38.38	0.886	0.885	0.895	0.822	0.812	54.55	60.61	47.47	60.61	56.57
Qwen3-4B	8.88	30.90	90.70	49.49	39.39	0.887	0.888	0.900	0.748	0.802	52.53	46.46	44.44	44.44	49.49
Qwen3-8B	5.65	31.71	89.89	41.41	50.00	0.879	0.873	0.892	0.754	0.774	54.55	56.57	47.47	49.49	55.56
Aya23-8B	8.28	23.03	80.00	28.78	14.14	0.896	0.894	0.902	0.803	0.786	57.58	47.47	49.49	57.58	46.46
Gemma2-9B-Instruct	9.09	39.19	91.51	81.81	49.49	0.903	0.898	0.908	0.879	0.812	60.61	42.42	50.51	56.57	39.39
Gemma3-4B-Instruct	23.43	33.93	90.50	62.62	38.38	0.872	0.857	0.875	0.827	0.764	41.41	42.42	44.44	57.58	34.34
Regional															
Sailor2-8B	27.67	40.40	86.46	-	-	0.736	0.660	0.588	-	-	45.45	49.49	31.31	-	-
Sahabat-AI-Gemma	13.73	43.23	93.53	-	-	0.899	0.895	0.907	-	-	62.63	53.54	44.44	-	-
SILMA-9B-Instruct	-	-	-	-	0.0	-	-	-	-	0.733	-	-	-	-	34.34
ALLAM-7B-Instruct	-	-	-	-	32.32	-	-	-	-	0.359	-	-	-	-	34.34

Table 6: **Experimental Results on PINGPONG.** This ablation is conducted under a zero-shot prompting setting, with the reasoning (thinking) mode disabled for models that support it. For the Question Answering task, we report results on the answerable subset.

tigate whether thinking trace is beneficial in our dataset. For models without built-in thinking behavior, we can elicit reasoning by explicitly asking the model to generate their reasoning trace first, akin to prior work such as chain-of-thought (Wei et al., 2022).

4.2 Task Setup

Across all downstream tasks, we adopt a set of shared experimental prompt configurations, which include the construction of example dialogues, the number of shots, and the specification of output formats. These configurations are further adapted to the requirements of each individual task. The prompts are detailed in Appendix D.

Number of Shots. For QA, we experiment with 0 and 1 shot prompting. For Dialogue Summarization, we use 0, 1, and 3-shot settings of summary examples without a dialogue example. The in-context examples for each task are drawn from dialogues that are not part of the evaluation set.

4.3 Evaluation Metrics

All models are assessed with task-specific evaluation metrics. For QUESTION ANSWERING and TOPIC CLASSIFICATION tasks, we use accuracy as the evaluation metric. For DIALOGUE SUMMARIZATION task, we use ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CHRF++ (Popović, 2015), and BERTScore (Zhang et al., 2020), as implemented in the Hugging Face `evaluate` library. These metrics were chosen because they have a high correlation with the four dimensions of summarization quality, according to (Fabbri et al., 2021). In the main paper we report BERTScore-F, but other metrics are shown in Appendix H.2.

5 Results and Analysis

The performance metrics for our evaluated models across all target languages are summarized in Table 6 (Full results are detailed in Appendix H). A key takeaway from these results is that the majority of models exhibit poor performance, which suggests that the proposed benchmark represents a significant challenges for current systems. Notably, we observe a a performance gap between general-purpose multilingual models and regionally-designed ones. The latter generally achieve better results, highlighting the clear benefits of utilizing specialized models that are tailored to specific linguistic and cultural characteristics.

While there is notable variance in performance across different language groups, it is difficult to conclude that any specific language is inherently more difficult than others. This is primarily due to the non-parallel nature of our dataset; the difficulty remains anchored to the specific conversation content of each language (see Appendix F and Appendix G).

5.1 Effect of Reasoning

We investigated the impact of explicit reasoning on model performance by utilizing the native "thinking traces" available in recent models like Qwen3, and by using explicit prompting to generate reasoning steps for other models. As shown in Figure 5, we observe a consistent improvement mostly in performance when reasoning is enabled. This indicates that the tasks within our dataset benefit from the additional computation and internal verification afforded by reasoning traces.

5.2 Effect of Few-Shot Learning

In comparison, the inclusion of few-shot examples does not appear to be a significant factor in

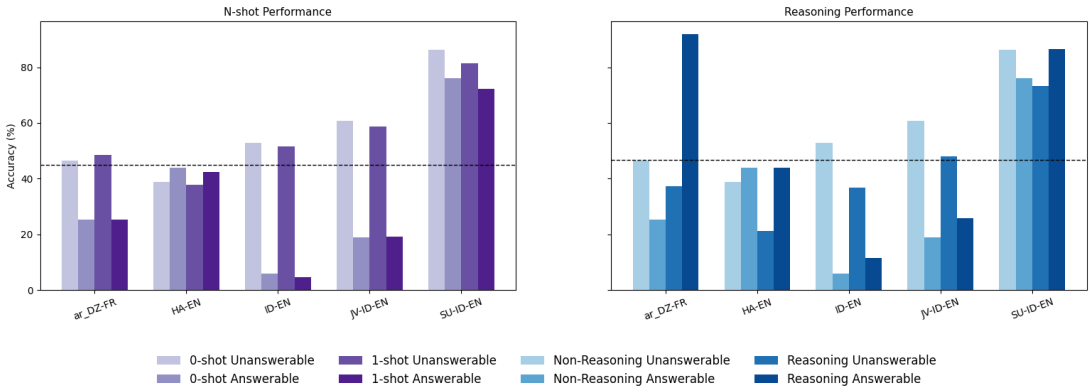


Figure 4: Comparison of average model performance (Acc. %) across languages for Answerable vs. Unanswerable cases, from the perspectives of N-shot prompting (Left) and reasoning performance (Right).

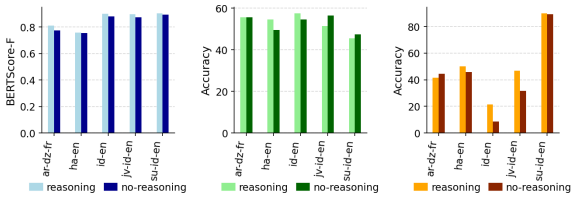


Figure 5: Reasoning vs No-Reasoning performance in Dialogue Summarization (blue), Topic Classification (green), and QA (red).

improving model performance for QA and Topic Classification, but improves slightly in Dialogue Summarization. As shown in Figure 6, providing a small number of in-context examples generally does not yield consistent gains across the evaluated models.

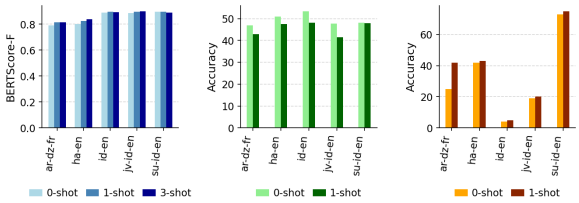


Figure 6: Zero-shot vs Few-shot performance in Dialogue Summarization (blue), Topic Classification (green), and QA (red).

5.3 Answerable vs. Unanswerable QA

Based on the results in Figure 4, we observe that answerable questions—specifically those designed to require explicit reasoning—exhibit consistent performance improvements when the model leverages reasoning mechanisms, whether implicitly or explicitly. In contrast, for the unanswerable subset, performance gains are observed only when explicit reasoning traces are enabled in the Qwen3 model, whereas implicit reasoning does not yield improvements and, in some cases, even leads to performance decline. This suggests that unanswerable cases place stricter demands on structured reasoning and error detection. Furthermore, a detailed analysis of shot settings indicates that few-shot prompting does not provide stable or uniform benefits across models. The effectiveness of few-shot examples appears to be model-dependent, with some models benefiting marginally while others show no improvement or even slight regressions.

6 Conclusion

In this work, we introduced PINGPONG, a multi-party code-switching benchmark capturing the authentic complexity of human multilingual discourse across five language combinations. By utilizing human-authored conversations with 2 to 4 participants, we provide a dataset containing structural nuances, such as multi-threaded dynamics and varied speaker dominance, often missing in synthetic corpora. Our analysis confirms that these dialogues are more natural and structurally diverse than machine-generated alternatives, particularly in message length and reply distances. We build an evaluation pipeline that covers three downstream tasks based on the conversation: QA, Summarization, and Topic Classification to address gaps in prior benchmarks and reflect the requirements of the modern LLM era. Experimental results show that state-of-the-art models still struggle with the intricacies of natural, multi-party code-switching. This performance gap highlights the importance of PINGPONG in identifying current NLP limitations and serves as a foundation for developing more robust, inclusive systems for the world's multilingual majority.

Limitations

Languages Coverage. This work covers 4 complementary (under-studied) languages across 5 language combinations, spanning 3 geographic regions. While not exhaustive, this scope provides a scalable foundation for building natural code-switching datasets such as PINGPONG, which can be readily extended to additional languages and regions in future work. As a future work, currently we are working on extending the coverage to cover more languagues, including Hindi-English combination. Due to practical constraints, the data cannot cover all regions and topics or reflect all cultural perspectives.

Models. We evaluate a diverse set of global and regional LLMs that vary in linguistic coverage, model size, and reasoning capability, including both open-source and proprietary systems. As the LLM landscape continues to evolve, expanding this evaluation to a broader range of models represents a natural and promising direction for future research.

Non-Parallel Dataset. Our dataset encompasses a wide range of code-switching language combinations, each annotated by native speakers of the corresponding language varieties under a unified annotation guideline. Consequently, the datasets and model performance measurements for each language setting are inherently independent rather than parallel, and therefore cannot be directly compared across languages. Nevertheless, this characteristic constitutes a key strength of our benchmark, as it preserves the naturalness and authenticity of the code-switching phenomena represented in the data.

Task Design. In the Answerable Question subset of the QA task, the questions are intentionally designed such that they can only be solved through reasoning augmented with external knowledge, particularly local or cultural knowledge specific to a given region. As a result, several language combinations exhibit substantial performance degradation, in some cases falling below random-chance performance (<50%). However, this property further demonstrates the challenging nature of our benchmark while simultaneously encouraging the development of fairer LLMs with more equitable representations of localized and culturally grounded knowledge across regions.

Evaluation Metric. Finally, due to the lack of reliable tools for computing advanced CMI metrics in under-studied languages, we adopt a relaxed CMI formulation that preserves meaningful comparative insights. We expect that future advances in multilingual NLP resources will enable more fine-grained code-mixing analyses for these languages.

Ethical Considerations

The annotators involved in this study were compensated above the local minimum wage. They received detailed instructions, and any demographic information collected was obtained with their informed consent (see Appendix A). Annotators were explicitly instructed not to use offensive language, and we conducted verification to the best of our ability. Nonetheless, some instances may have been missed, and we are committed to updating the benchmark if such cases are reported. While we aimed to construct realistic datasets, we do not claim that our benchmark captures all code-switching variations across the five language combinations. We also acknowledge the use of an AI assistant (e.g. ChatGPT, Gemini, Github Copilot) to support code implementation and to polish the writing of this manuscript.

References

Mofetoluwa Adeyemi, Akintunde Oladipo, Xinyu Zhang, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Boxing Chen, and Jimmy Lin. 2023. Ciral at fire 2023: Cross-lingual information retrieval for african languages. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 4–6.

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, Genta Indra Winata, Pascale Fung, and Ayu Purwarianti. 2022. IndoRobusta: Towards robustness against diverse code-mixed Indonesian local languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 25–34, Online. Association for Computational Linguistics.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named entity recognition on code-switched data: Overview of the calcs 2018 shared task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020a. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the*

Twelfth Language Resources and Evaluation Conference, pages 1803–1813, Marseille, France. European Language Resources Association.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020b. Lince: A centralized benchmark for linguistic code-switching evaluation. *arXiv preprint arXiv:2005.04322*.

Gustavo Aguilar and Thamar Solorio. 2020. From english to code-switching: Transfer learning with strong morphological clues. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8033–8044.

Maia Aguirre, Laura García-Sardiña, Manex Serras, Ariane Méndez, and Jacobo López. 2022. BaSCo: An annotated Basque-Spanish code-switching corpus for natural language understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3158–3163, Marseille, France. European Language Resources Association.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, and 1 others. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Peter Auer. 2013. *Code-switching in conversation: Language, interaction and identity*. Routledge.

Oreoluwa Boluwatife Babatunde, Victor Tolulope Olufemi, Emmanuel Bolarinwa, Kausar Yetunde Moshood, and Chris Chinenye Emezue. 2025. Beyond monolingual limits: Fine-tuning monolingual asr for yoruba-english code-switching. In *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 18–25.

Mohamed Balabel, Injy Hamed, Slim Abdennadher, Ngoc Thang Vu, and Özlem Çetinoğlu. 2020. Cairo student code-switch (CSCS) corpus: An annotated Egyptian Arabic-English corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3973–3977, Marseille, France. European Language Resources Association.

Kelsey Ball and Dan Garrette. 2018. Part-of-speech tagging for code-switched, transliterated texts without explicit language identification. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3084–3089.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. ALLam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Barbara E Bullock and Almeida Jacqueline Toribio. 2009. *The Cambridge handbook of linguistic code-switching*. Cambridge university press.

Laurie Burchell, Alexandra Birch, Robert Thompson, and Kenneth Heafield. 2024. Code-switched language identification is harder than you think. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 646–658.

Özlem Çetinoğlu and Çağrı Çöltekin. 2016. Part of speech annotation of a turkish-german code-switching corpus. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 120–130.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P. McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Richeek Das, Sahasra Ranjan, Shreya Pathak, and Preethi Jyothi. 2023. Improving pretraining techniques for code-switched nlp. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1176–1191.

Longxu Dou, Qian Liu, Fan Zhou, Changyu Chen, Zili Wang, Ziqi Jin, Zichen Liu, Tongyao Zhu, Cunxiao Du, Penghui Yang, and 1 others. 2025. Sailor2: Sailing in south-east asia with inclusive multilingual llms. *arXiv preprint arXiv:2502.12982*.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855.

Ayush Garg, Sammed Kagi, Vivek Srivastava, and Mayank Singh. 2021. MIPE: A metric independent pipeline for effective code-mixed NLG evaluation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 123–132, Punta Cana, Dominican Republic. Association for Computational Linguistics.

François Grosjean. 2010. *Bilingual: Life and reality*. Harvard university press.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. ArzEn: A speech corpus for code-switched Egyptian Arabic-English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4237–4246, Marseille, France. European Language Resources Association.

Maite Heredia, Jeremy Barnes, and Aitor Soroa. 2025. Euska\˜ nolds: A naturally sourced corpus for basque-spanish code-switching. *arXiv preprint arXiv:2502.03188*.

Megan Herrera, Ankit Aich, and Natalie Parde. 2022. TweetTaglish: A dataset for investigating Tagalog-English code-switching. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2090–2097, Marseille, France. European Language Resources Association.

Devanshu Jain, Maria Kustikova, Mayank Darbari, Rishabh Gupta, and Stephen Mayhew. 2018. Simple features for strong performance on named entity recognition in code-switched twitter data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 103–109.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020a. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020b. Gluecos: An evaluation benchmark for code-switched nlp. *arXiv preprint arXiv:2004.12376*.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Nikita Kulkarni, Kareena Manghani, Sanhita Kulkarni, Pranita Deshmukh, and Raviraj Joshi. 2024. L3cube-mahasum: A comprehensive dataset and bart models for abstractive text summarization in marathi. In *Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 76–79.

Garry Kuwanto, Chaitanya Agarwal, Genta Indra Winata, and Derry Tanti Wijaya. 2024. Linguistics theory meets llm: Code-switched text generation via equivalence constrained large language models. *arXiv preprint arXiv:2410.22660*.

Ying Li and Pascale Fung. 2012. Code-switch language model with inversion constraints for mixed language speech recognition. In *Proceedings of COLING 2012*, pages 1671–1680.

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

Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Peng Xu, Yan Xu, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J Barezi, and 1 others. 2022. Ascend: A spontaneous chinese-english dataset for code-switching in multi-turn conversation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7259–7268.

Dau-Cheng Lyu, Tien Ping Tan, Engsiong Chng, and Haizhou Li. 2010. Seame: a mandarin-english code-switching speech corpus in south-east asia. In *Interspeech*, volume 10, pages 1986–1989.

Piyush Makhija, Ankit Kumar, and Anuj Gupta. 2020. hinglishNorm - a corpus of Hindi-English code mixed sentences for text normalization. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 136–145, Online. International Committee on Computational Linguistics.

Thomas Mandl, Koyel Ghosh, Nishat Raihan, Sandip Modha, Shrey Satapara, Tanishka Gaur, Yaashu Dave, Marcos Zampieri, and Sylvia Jaki. 2024. Overview of the hasoc track 2024: Hate-speech identification in english and bengali. In *Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 1–2.

Laiba Mehnaz, Debanjan Mahata, Rakesh Gosangi, Uma Sushmitha Gunturi, Riya Jain, Gauri Gupta, Amardeep Kumar, Isabelle G. Lee, Anish Acharya, and Rajiv Ratn Shah. 2021. GupShup: Summarizing open-domain code-switched conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6177–6192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

Li Nguyen and Christopher Bryant. 2020. CanVEC – the Canberra Vietnamese-English code-switching natural speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4121–4129, Marseille, France. European Language Resources Association.

Jessica Ojo, Zina Kamel, and David Ifeoluwa Adelani. 2025. Divers-bench: Evaluating language identification across domain shifts and code-switching. *arXiv preprint arXiv:2509.17768*.

Tanmay Parekh, Emily Ahn, Yulia Tsvetkov, and Alan W Black. 2020. Understanding linguistic accommodation in code-switched human-machine dialogues. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 565–577, Online. Association for Computational Linguistics.

Shana Poplack. 2013. "sometimes i'll start a sentence in spanish y termino en español": Toward a typology of code-switching. *Linguistics*, 51(s1):11–14.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Justus Randolph. 2010. Free-marginal multirater kappa (multirater free): An alternative to fleiss fixed-marginal multirater kappa. volume 4.

Priya Rani, John P. McCrae, and Theodorus Fransen. 2022. MHE: Code-mixed corpora for similar language identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3425–3433, Marseille, France. European Language Resources Association.

Manikandan Ravikiran and Subbiah Annamalai. 2021. DOSA: Dravidian code-mixed offensive span identification dataset. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 10–17, Kyiv. Association for Computational Linguistics.

Yurii Shynkarov, Veronika Solopova, and Vera Schmitt. 2025. Improving sentiment analysis for ukrainian social media code-switching data. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 179–193.

silma-ai. 2024. Silma 9b instruct v1.0. `https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0`.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Victor Soto and Julia Hirschberg. 2018. Joint part-of-speech and language id tagging for code-switched data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–10.

Vivek Srivastava and Mayank Singh. 2020. PHINC: A parallel Hinglish social media code-mixed corpus for machine translation. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 41–49, Online. Association for Computational Linguistics.

Vivek Srivastava and Mayank Singh. 2021. HinGE: A dataset for generation and evaluation of code-mixed Hinglish text. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 200–208, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sathya Krishnan Suresh, Tanmay Surana, Lim Zhi Hao, and Eng Siong Chng. 2025. Cs-sum: A benchmark for code-switching dialogue summarization and the limits of large language models. *arXiv preprint arXiv:2505.13559*.

Ishan Tarunesh, Syamantak Kumar, and Preethi Jyothi. 2021. From machine translation to code-switching: Generating high-quality code-switched text. 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 3154–3169, Online. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin,

Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

G Richard Tucker. 1999. A global perspective on bilingualism and bilingual education. *Washington, DC*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Li Wei. 2018. Translanguaging as a practical theory of language. *Applied linguistics*, 39(1):9–30.

Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978.

Genta Indra Winata, David Anugraha, Patrick Amadeus Irawan, Anirban Das, Haneul Yoo, Paresh Dashore, Shreyas Kulkarni, Ruochen Zhang, Haruki Sakajo, Frederikus Hudi, and 1 others. 2026. Can large language models understand, reason about, and generate code-switched text? *arXiv preprint arXiv:2601.07153*.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153.

Genta Indra Winata, Zhaojiang Lin, and Pascale Fung. 2019a. Learning multilingual meta-embeddings for code-switching named entity recognition. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 181–186.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Code-switching language modeling using syntax-aware multi-task learning. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 62–67.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019b. Code-switched language models using neural based synthetic data from parallel sentences. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 271–280.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zheng-Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Fikri Aji. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics.

Ruochen Zhang and Carsten Eickhoff. 2024. CroCoSum: A benchmark dataset for cross-lingual code-switched summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4113–4126, Torino, Italia. ELRA and ICCL.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 9220–9230.

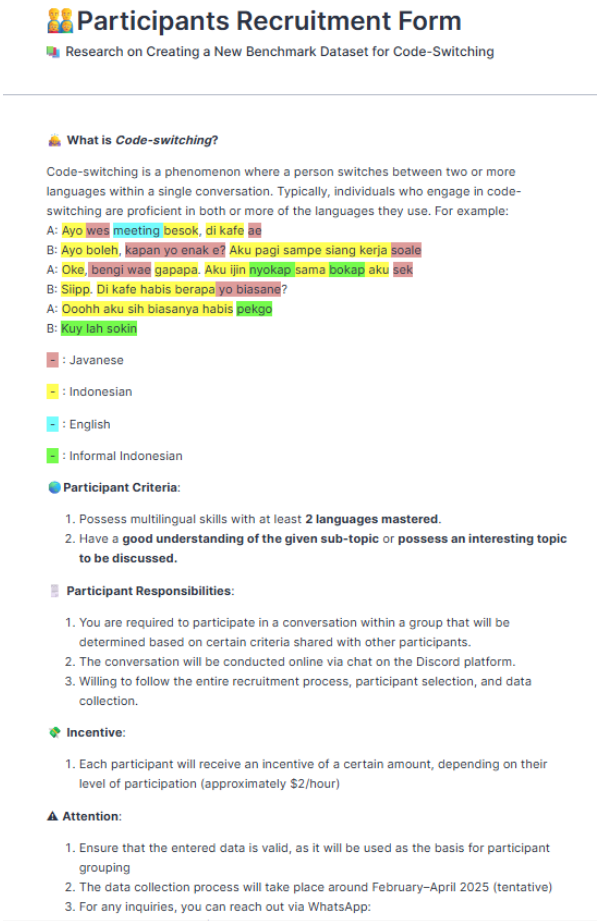
A Recruitment Form

The recruitment form is organized into three main sections, as described below.

Introduction. This section provides an overview and general background on code-switching, along with the eligibility criteria required for respondents to participate in the crowdsourcing process. The introduction section of the recruitment form is shown in Figure 7. This also shows what kind of data that will be collected from the participant in the process.

Self-Assessment. Respondents are asked to provide personal demographic information, as well as details about the languages they commonly use when communicating within their close social circles. The self-assessment section of the recruitment form is illustrated in Figures 8, Figure 9, and Figure 10.

Language Assessment. The recruitment form also includes a language assessment component, in which respondents are instructed to compose a short paragraph on a given topic using a specified combination of languages. This task is designed to simulate respondents' natural code-switching behavior. The language assessment section of the recruitment form is shown in Figure 11.



Participants Recruitment Form

Research on Creating a New Benchmark Dataset for Code-Switching

What is *Code-switching*?

Code-switching is a phenomenon where a person switches between two or more languages within a single conversation. Typically, individuals who engage in code-switching are proficient in both or more of the languages they use. For example:

A: Ayo wes meeting besok, di kafe ae
B: Ayo boleh, kapan yo enak e? Aku pagi sampe siang kerja soale
A: Oke, bengi wae gapapa. Aku ijin nyokap sama bokap aku sek
B: Siipp. Di kafe habis berapa yo biasane?
A: Ooohh aku sih biasanya habis pekgo
B: Kuy lah sokin

- : Javanese
- : Indonesian
- : English
- : Informal Indonesian

Participant Criteria:

1. Possess multilingual skills with at least **2 languages mastered**.
2. Have a **good understanding of the given sub-topic** or **possess an interesting topic to be discussed.**

Participant Responsibilities:

1. You are required to participate in a conversation within a group that will be determined based on certain criteria shared with other participants.
2. The conversation will be conducted online via chat on the Discord platform.
3. Willing to follow the entire recruitment process, participant selection, and data collection.

Incentive:

1. Each participant will receive an incentive of a certain amount, depending on their level of participation (approximately $2/hour)

Attention:

1. Ensure that the entered data is valid, as it will be used as the basis for participant grouping
2. The data collection process will take place around February–April 2025 (tentative)
3. For any inquiries, you can reach out via WhatsApp:

Figure 7: Introduction Section of the Recruitment Form.

B Annotator Demographics

Table 7 presents the demographic data of the annotators involved in the dataset creation.

		Gender		Age (Years)				
Lang	**#Annotators**	**Male**	**Female**	**18-24**	**25-34**	**35-44**	**45-54**	**55+**
ID-EN	11	11	0	11	0	0	0	0
JV-ID-EN	30	10	20	26	4	0	0	0
SU-ID-EN	13	6	7	8	5	0	0	0
HA-EN	11	6	5	4	1	4	2	0
AR-DZ-FR	11	7	4	1	9	0	0	1

Table 7: Annotator Demographics Data

C Annotator Guideline

For each annotator-related task, a set of guidelines is provided to ensure that dataset construction is carried out consistently across languages.

Dialogue Construction. Guidelines for dialogue construction are provided through a dedicated channel on the Discord platform, which is used for collecting the dialogue data. These guidelines describe the participants' tasks, explain the grouping process, and specify the rules that must be followed during dialogue creation. The dialogue construction guidelines are shown in Figures 12 and 13.

Question Answering. Guidelines for the Question Answering task are presented on the landing page of the annotation platform. These guidelines outline the annotators' responsibilities, the types of questions that must be created, and examples of appropriate questions. The Question Answering guidelines are shown in Figure 14.

Dialogue Summarization. Guidelines for Dialogue Summarization are also provided on the landing page of the annotation platform. They describe the annotators' tasks, as well as general criteria for producing high-quality summaries, accompanied by illustrative examples. The guidelines for this task are shown in Figure 15.

Naturalness. The Naturalness guidelines are presented in the same manner as those for Question Answering and Dialogue Summarization. This section specifies the annotators' tasks and details the scoring criteria, which are based on prior work by

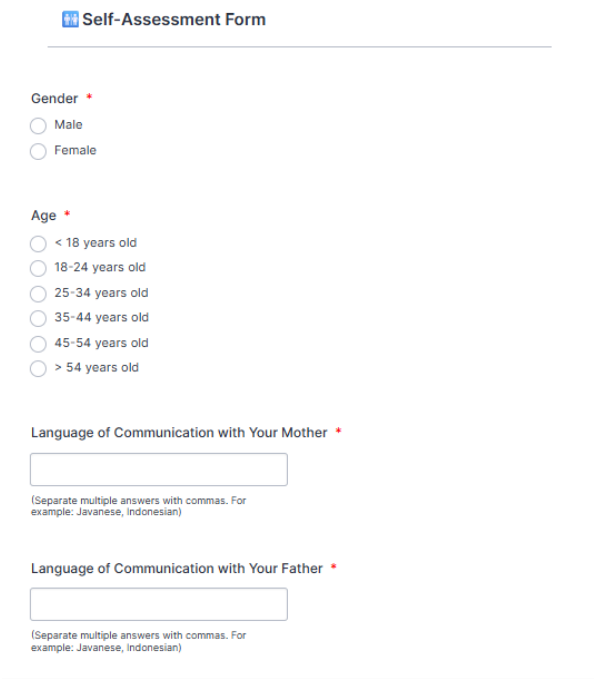


Figure 8: Self-Assessment Section (Part 1) of the Recruitment Form.

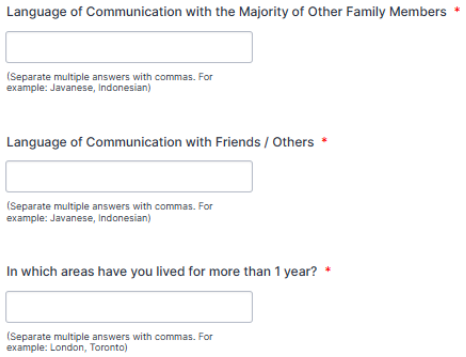


Figure 9: Self-Assessment Section (Part 2) of the Recruitment Form.

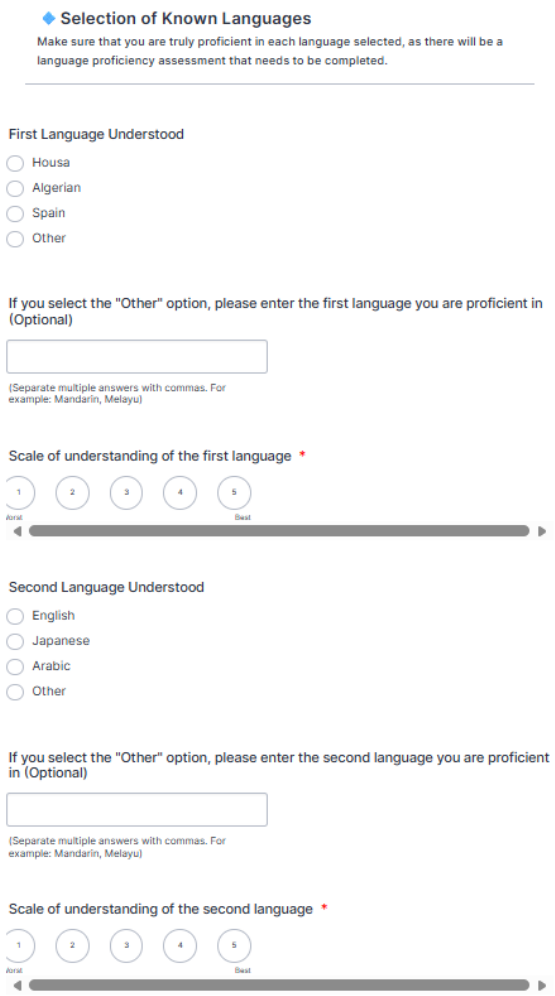


Figure 10: Self-Assessment Section (Part 3) of the Recruitment Form.

Yong et al. (2023). The Naturalness guidelines are shown in Figure 16.

D Prompt Example

The prompts used in our experiments consist of four components, as described below.

Question Answering. Figure 17 illustrates an example of the prompt template used for inference in the Question Answering (QA) task. In the figure, the section highlighted in yellow is included only when the few-shot prompting option is enabled.

Dialogue Summarization. Figure 18 presents an example of the prompt template used for inference in the Dialogue Summarization task. Similar to the QA setting, the yellow-highlighted section is included only when the few-shot prompting option is applied.

Topic Classification. Figure 19 illustrates an example of the prompt used for the Topic Classification task. This prompt uses one-shot setting and instructs the model to provide an explanation for its predicted label.

Machine-Generated Conversation. Figure 20 illustrates an example of the prompt template used for inference in the Machine-Generated Conversation task. This prompt is used to perform inference with the GPT-4o model, enabling it to generate dialogue text that adheres to the specifications defined in the prompt.

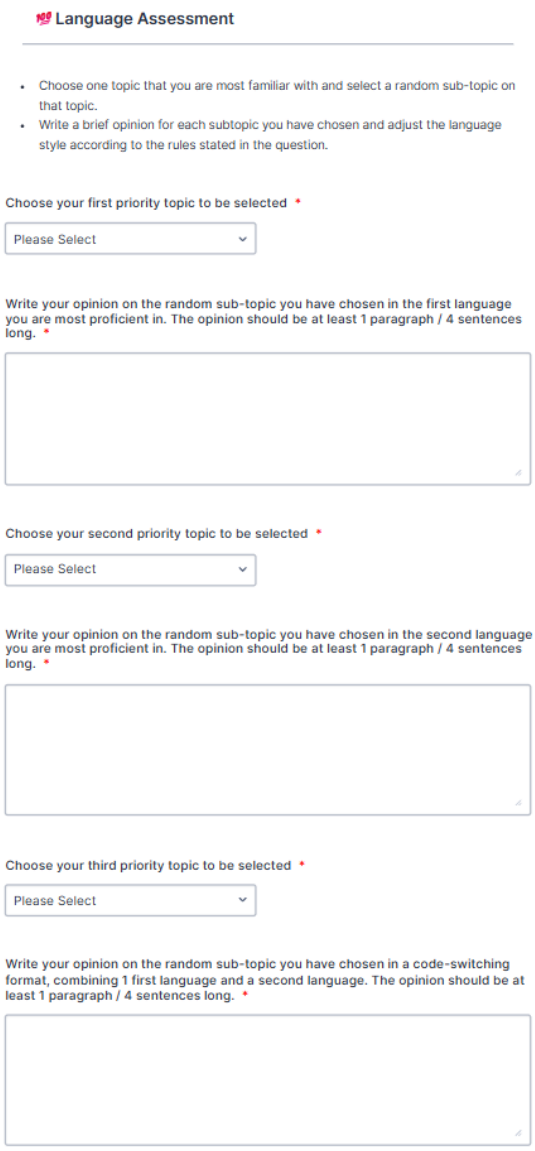


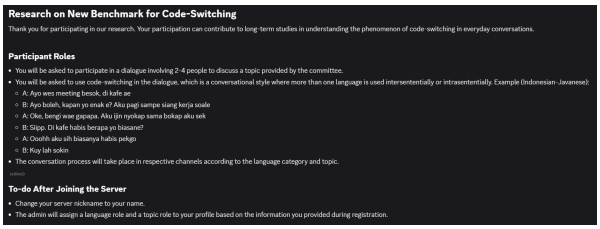
Figure 11: Language-Assessment Section of the Recruitment Form.

E Hyperparameter Settings

This section documents the hyperparameter configurations adopted for each model across the three downstream tasks. A complete summary of the settings is provided in Table 8. Given the substantial computational costs, all evaluations were limited to a single run. Because our core novelty lies in the dataset itself, our objective is not to provide an absolute model leaderboard, but rather to establish baseline performance across a diverse range of models and configurations.

F Qualitative Analysis of QA Topic

Based on the ablation results presented in Table 6, several language configurations consistently exhibit substantially lower performance, often falling below random-chance level (<50%), particularly Indonesian-English (ID-EN) and Javanese-Indonesian-English (JV-ID-EN). We hypothesize that this degradation primarily stems from the non-parallel nature of our dataset construction. Unlike parallel multilingual benchmarks, each language



Research on New Benchmark for Code-Switching

Thank you for participating in our research. Your participation can contribute to long-term studies in understanding the phenomenon of code-switching in everyday conversations.

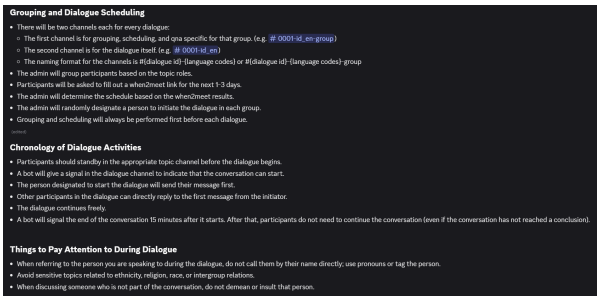
Participant Roles

- You will be asked to participate in a dialogue involving 2-4 people to discuss a topic provided by the committee.
- You will be asked to use code-switching in the dialogue, which is a conversational style where more than one language is used intersententially or intrasententially. Example (Indonesian-Javanese):
  - A: Ayo wes meeting besok, di kafe ae
  - B: Ayo boleh, kapan yo enak ti? Aku pagi sampe siang kerja soale
  - A: Oke, bengi wae gapapa. Aku ijin nyokap sama bokap aku sek
  - B: Siipp. Di kafe habis berapa yo biasane?
  - A: Ooohh aku sih biasanya habis pekgo
  - B: Kuy lah sokin
- The conversation process will take place in respective channels according to the language category and topic.

To-do After Joining the Server

- Change your server nickname to your name.
- The admin will assign a language role and a topic role to your profile based on the information you provided during registration.

Figure 12: Dialog Construction Annotator Guideline (Part 1).



Grouping and Dialogue Scheduling

- There will be two channels each for every dialogue:
  - The first channel is for grouping, scheduling, and qna specific for that group. (e.g. # 0001-id_en-group)
  - The second channel is for the dialogue itself. (e.g. # 0001-id_en)
  - The naming format for the channels is #{dialogue id}-{language codes} or #{dialogue id}-{language codes}-group
- The admin will group participants based on the topic roles.
- Participants will be asked to fill out a when2meet link for the next 1-3 days.
- The admin will determine the schedule based on the when2meet results.
- The admin will randomly designate a person to initiate the dialogue in each group.
- Grouping and scheduling will always be performed first before each dialogue.

Chronology of Dialogue Activities

- Participants should standby in the appropriate topic channel before the dialogue begins.
- A bot will give a signal in the dialogue channel to indicate that the conversation can start.
- The person designated to start the dialogue will send their message first.
- Other participants in the dialogue can directly reply to the first message from the initiator.
- The dialogue continues freely.
- A bot will signal the end of the conversation 15 minutes after it starts. After that, participants do not need to continue the conversation (even if the conversation has not reached a conclusion).

Things to Pay Attention to During Dialogue

- When referring to the person you are speaking to during the dialogue, do not call them by their name directly; use pronouns or tag the person.
- Avoid sensitive topics related to ethnicity, religion, race, or intergroup relations.
- When discussing someone who is not part of the conversation, do not demean or insult that person.

Figure 13: Dialog Construction Annotator Guideline (Part 2).

Model	Temperature	Top-P	Top-K	Min-P	Max-Tokens
Question Answering					
Qwen2.5-3B-Instruct	0.7	0.8	20	0	512
Qwen2.5-7B-Instruct	0.7	0.8	20	0	512
Qwen3-4B					
Think	0.6	0.95	0.20	0	8192
No-Think	0.7	0.8	20	0	8192
Qwen3-8B					
Think	0.6	0.95	0.20	0	8192
No-Think	0.7	0.8	20	0	8192
Aya23-8B	0.3	1.0	1	0	512
Gemma2-9B-Instruct	0.1	0.9	5	0	512
Gemma3-4B-Instruct	0.2	0.9	10	0	512
Sailor2-8B	0.3	0.9	30	0	512
Sahabat-AI-Gemma	0.1	0.9	5	0	512
SILMA-9B-Instruct	0.3	0.9	10	0	512
ALLAM-7B-Instruct	0.4	0.9	20	0	512
Text Summarization & Topic Classification					
Qwen2.5-3B-Instruct	0.4	0.8	20	0	2048
Qwen2.5-7B-Instruct	0.4	0.8	20	0	2048
Qwen3-4B					
Think	0.6	0.95	20	0	8192
No-Think	0.7	0.8	20	0	2048
Qwen3-8B					
Think	0.6	0.95	20	0	8192
No-Think	0.7	0.8	20	0	2048
Aya23-8B	0.3	0.8	20	0	2048
Gemma2-9B-Instruct	0.4	0.8	20	0	2048
Gemma3-4B-Instruct	0.4	0.8	20	0	2048
Sailor2-8B-Chat	0.4	0.8	20	0	2048
Sahabat-AI-Gemma	0.4	0.8	20	0	2048
SILMA-9B-Instruct	0.4	0.8	20	0	2048
ALLAM-7B-Instruct	0.6	0.95	50	0	2048

Table 8: Hyperparameter settings used for each model across all tasks.

Research on Creating a New Benchmark for Code-Switching

Thank you for participating in our research. Your participation has contributed to the creation of a new code-switching benchmark and has helped facilitate our research process.

Criteria:

- Possesses multilingual proficiency in the following languages: Sundanese, Indonesian, English
- Has the ability to think critically when formulating questions

Task:

- You are required to create several *multiple-choice* questions based on specific criteria.
- There are two categories of questions that can be created: **Answerable** and **Unanswerable**.
- Each **multiple-choice question** should be convertible into a **short question-answer** format.

Example
1. How old is Speaker A?
2. Why does Speaker A want to go to Bandung?
3. What is Speaker C's grandmother's favorite color?

Criteria for **Answerable** question:

- When creating **Answerable** questions, each question **MUST REQUIRE REASONING** to answer, such that answers could not be obtained directly from the dialogue. Instead, they required the use of external knowledge and structured reasoning to answer the question.
- For each question, provide **1 correct answer, 3 distractor options, and 1 option "No correct answer."** This will result in answer choices A, B, C, D, and E for each question.

Criteria for **Unanswerable** question:

- When creating **Unanswerable** questions, there are 5 types of questions that can be created:
1. **Negation**: Add or remove a word with a negation meaning in the dialogue.
   Example: If the dialogue discusses Speaker A planning to move to a new boarding house:
   Question: "Why does Speaker A not intend to move to a new boarding house?"
2. **Antonym**: Use an antonym for a specific word in the dialogue.
   Example: If the dialogue discusses the Indonesian national football team's performance improving:
   Question: "What caused the Indonesian national football team's performance to worsen?"
3. **Entity swap**: Replace entities, numbers, or dates with other entities, numbers, or dates.
   Example: If the dialogue discusses a dinner event on September 9, 2025:
   Question: "How far has the preparation for the dinner event on January 10, 2024, progressed?"
4. **Mutual exclusion**: The question contradicts a statement in the dialogue.
   Example: If the dialogue discusses the government's focus on developing environmentally friendly electric vehicle technology:
   Question: "How is the government working to advance the development of solar-powered car technology?"
5. **Impossible condition**: The question cannot be answered at all with the information in the dialogue, but is still related to the context of the conversation.
   Example: If the dialogue is about Indonesian political developments after the local elections:
   Question: "Why did one of the gubernatorial candidates in the recent local elections make so many blunders?"

- For each question, provide **4 distractor options and 1 option "No correct answer."** This will result in answer choices A, B, C, D, and E for each question.

Instructions for Work:

1. Choose the dialogue to be used (prioritize dialogues at the top of the list that have not yet had their questions created) through the following link:
<LINK TO DIALOG LIST>
(Note: Choose a dialogue in a language you understand)
2. Download the selected dialogue and carefully read its contents along with the topics involved.
3. The allotted time for creating questions is 15 minutes with the following conditions:
- For Unanswerable questions, you are free to choose the question type (the types must be different for each question).
4. Write each question in the provided column in the following format:
Example:
What is the difference between the selling price and the buying price of item A?
a. 100 thousand rupiahs
b. Rp3000
c. 6K rupiahs
d. 30,000 rupiahs
e. No correct answer
5. Write the correct answer key by only writing its letter in the provided column.
Example:
a

Figure 14: Question Answering Annotator Guideline.

Construction of a Code-Switched Dialogue Summarization Dataset

Thank you for participating in our research. Your participation helps support long-term studies aimed at understanding code-switching phenomena in everyday conversations.

What is code-switching?

Code-switching is a phenomenon in which a speaker alternates between two or more languages within a single conversation. Individuals who engage in code-switching typically possess proficiency in all languages involved. For example:
A: Ayo wes meeting besok, di kafe ae
B: Ayo boleh, kapan yo enak e? Aku pagi sampe siang kerja soale
A: Oke, bengi wae gapapa. Aku jln nyokap sama bokap aku sek
B: Siipp. Di kafe habis berapa yo biasane?
A: Ooohh aku sih biasanya habis pekgo
B: Kuy lah sokin

Your Task as a Participant

- You are asked to write a summary of a dialogue that contains code-switching.
- The summary should consist of 3–5 sentences written in formal Indonesian.
- Each dialogue is accompanied by a specified topic. The summary must be constrained to this topic; any information that is not relevant to the given topic should be excluded from the summary (this relates to the relevance aspect of summarization).
- In the written summary, you may refer to speakers in the dialogue directly using labels such as A, B, C, and so on, as indicated in the dialogue.
- Several aspects related to summary quality should be considered when writing the summary, which will be explained in the following section.

Aspects of Summary Quality

Fluency

The quality of each individual sentence in the summary. Sentences in the summary should be free from formatting errors, capitalization errors, and grammatical mistakes that may hinder readability.

Example of a well-formed sentence	Example of a bad-formed sentence
Negara dengan populasi keempat terbesar di dunia adalah Indonesia.	negara dgn poplasi keempat Terbesar didunia adalah indonesia.

Consistency

Factual consistency between the summary and the source dialogue. A factually consistent summary contains only statements that are supported by the source document. Avoid producing summaries that include statements not grounded in the source (hallucinations).

Relevance

Selection of important content from the source. The summary should include only essential information related to the main topic of the source document. Avoid summaries that contain irrelevant or redundant information.

Coherence

The collective quality of all sentences in the summary. The summary should be well structured and clearly organized. It should not merely consist of a collection of related statements, but should develop coherently from sentence to sentence into a unified body of information about the topic, with clear logical connections across sentences.

You may use appropriate conjunctions and word choices to ensure paragraph-level coherence.

Example of a well-formed paragraph	Example of a bad-formed paragraph
Industri ini memiliki banyak keunggulan. Industri ini dapat menarik banyak perhatian terhadap riset, terutama dalam hal strategi yang digunakan untuk memasuki pasar baru. Meskipun ada potensi pertumbuhan yang signifikan dalam jangka menengah hingga panjang, jelas bahwa pekerja di sektor ini membutuhkan pelatihan. Oleh karena itu, penting untuk menjaga agar staf tetap terupdate dengan perangkat lunak yang digunakan dalam industri ini.	Industri ini memiliki banyak keunggulan. Industri ini dapat menarik banyak perhatian untuk riset yang terkait. Industri ini juga dapat mengekspor produk ke mitra dagang utama. Industri ini memiliki potensi pertumbuhan yang signifikan dalam jangka menengah hingga panjang. Pekerja di industri ini membutuhkan pelatihan. Penting untuk menjaga agar staf tetap terupdate dengan perangkat lunak yang digunakan dalam industri ini.

Summary Writing Guidelines

- Please open the spreadsheet that has been assigned to you.
- There are four columns, defined as follows:
  - Dialogue: Dialog yang harus dirangkum.Dialogue: the dialogue to be summarized
  - Summary: The column for entering the summary.
  - Initial Topic: The topic associated with the dialogue.
  - Topic Familiarity: Check the box if you feel sufficiently familiar with the topic of the conversation.
- Please complete the summary in the Summary column.

Figure 15: Dialogue Summarization Annotator Guideline.

configuration in our dataset contains distinct forms of local knowledge embedded within the dialogue context, alongside varying conversational topics and culturally grounded references across language pairs.

To further investigate this phenomenon, Table 9 presents a qualitative analysis of the two lowest-performing language configurations. For each language pair, we selected the top-10 and bottom-10 dialogues based on the average performance of all evaluated LLMs. Human annotators were then asked to assess whether the dialogue topics were excessively specific, such as containing strong local knowledge dependencies, references tied to the original annotator's personal experiences, or temporally constrained contexts that may not generalize across linguistic settings.

Construction of a Code-Switched Dialogue Summarization Dataset

Thank you for participating in our research. Your contribution helps support long-term research aimed at better understanding code-switching phenomena in everyday conversations.

What is code-switching?

Code-switching is a phenomenon in which a speaker alternates between two or more languages within a single conversation. Individuals who engage in code-switching typically possess proficiency in all languages involved. For example:
A: Ayo wes meeting besok, di kafe ae
B: Ayo boleh, kapan yo enak e? Aku pagi sampe siang kerja soale
A: Oke, bengi wae gapapa. Aku jln nyokap sama bokap aku sek
B: Siipp. Di kafe habis berapa yo biasane?
A: Ooohh aku sih biasanya habis pekgo
B: Kuy lah sokin

Your Task as a Participant

- You are asked to provide an assessment of the **naturalness** of the given dialogue.

Naturalness Evaluation Guidelines

- Please open the WorkingSpace tab, which contains several columns.
- There are six columns, defined as follows:
  - Dialogue Topic: the topic discussed in the dialogue
  - Language: the language(s) used by the speakers in the dialogue
  - Dialogue A: the first dialogue corresponding to the language and topic in that row
  - Score A: the field where you assign a naturalness score to Dialogue A
  - Dialogue B: the second dialogue corresponding to the language and topic in that row
  - Score B: the field where you assign a naturalness score to Dialogue B
- Please assign a numerical score to each dialogue according to the criteria above (e.g., 1, 2, etc.).

Naturalness Score

- Please follow the scoring scheme below when assigning a naturalness score to each dialogue:
  - 1: Not natural at all.
  - 2: The dialogue appears to be produced by speakers from a different region or cultural background (i.e., not expressed as by a native speaker of the language).
  - 3: As a native speaker of the language, the dialogue reflects how you yourself would naturally communicate.
- When assigning scores, ensure that each evaluation is made with respect to the language used in the dialogue and adheres strictly to the criteria above.

Figure 16: Human-written vs Machine-generated Naturalness Annotator Guideline.

The analysis reveals a clear trend. In the ID-EN configuration, 7 out of the 10 lowest-performing dialogues were categorized as overly specific. A similar pattern was observed in JV-ID-EN, where 8 out of the 10 bottom-ranked dialogues contained highly specific contextual content. In contrast, among the top-10 dialogues with the highest average model performance, only 2 dialogues in both ID-EN and JV-ID-EN were considered excessively specific. These findings provide supporting evidence for our hypothesis that the non-parallel structure of the dataset introduces substantial contextual variability across language configurations.

G Qualitative Analysis of Summarization

Beyond the quantitative metrics, we examine specific failure patterns observed in the summarization outputs to better understand model behavior across languages and prompting conditions.

G.1 Few-Shot Context Overflow

A notable failure was observed for **Sailor2-8B** and **ALLAM-7B-Instruct** under few-shot settings. Both models have a maximum context length of 4,096 tokens, and when few-shot examples are included in the prompt, the total input length frequently exceeds this limit. As a result, the API returns an HTTP 400 error of the following form:

```
"This model's maximum context length is
4096 tokens. However, you requested 2048
output tokens and your prompt contains
at least 2049 input tokens, for a total
of at least 4097 tokens."
```

This explains the dramatic BERTScore collapse observed for both models at 1-shot and 3-shot settings - the outputs are not degenerate summaries, but rather missing entirely due to failed API calls.

	ID-EN				JV-ID-EN			
	Avg. (%)	ID	Topic	Too Specific (?)	Avg. (%)	ID	Topic	Too Specific (?)
Top-10								
	47.78	19	All about BlackPink	✗	60.00	22	Experiences of being admitted to medical school	✓
	47.78	69	Talking about the latest chatbots/LLMs	✗	60.00	88	Discussions on patents and copyright protection	✗
	45.56	17	All about bioinformatics	✗	58.89	25	The lack of digital literacy interest among Indonesian youth	✗
	44.44	16	All about anime, favorites or recently watched	✗	55.56	90	Tourism villages and local community development	✗
	44.44	75	Discussion on ISAE-SUPAERO's master's programs	✓	54.44	13	Experiences of participating in debate competitions during high school	✓
	43.33	20	Recently played games	✓	54.44	46	Interest in learning foreign languages	✗
	42.22	2	Current developments in electric vehicle technology	✗	53.33	24	Perspectives on the legalization of cannabis for medical purposes	✗
	41.11	76	Talking about various renewable energy sources	✗	53.33	53	Being a morning person versus a night owl	✗
	40.00	68	Discussing the Israel-Hamas ceasefire	✗	52.22	2	Opinions on Manchester United	✗
	40.00	81	Discussing Trump's new policies	✗	52.22	15	Experiences of studying in the US	✗
Bottom-10								
	10.00	92	Discussing various music platforms	✗	17.78	40	Perspectives on the COC event by Ruangguru	✓
	13.33	99	Public transportation in Bandung	✓	20.00	61	Experiences of participating in "mlaku-mlaku" activities	✓
	14.44	95	Complaints about Edunex	✓	24.44	8	Perspectives on the inauguration of Indonesia's new president in 2024	✗
	15.56	79	Pros and cons of ITB's fast-track program	✓	24.44	34	Weather predictions for Indonesian cities based recent days	✓
	16.67	15	Current meta heroes in Mobile Legends	✓	24.44	76	Experiences of falling before a proposal seminar presentation	✓
	16.67	18	Thoughts on Apple's M4 chip	✗	26.67	70	Discussions about proposal seminars (seminar proposal / sempro)	✓
	16.67	61	Thoughts on Indonesia joining BRICS	✗	27.78	3	Experiences related to local culinary traditions	✓
	17.78	22	Recent activities	✓	27.78	35	Favorite music genres and musicians	✓
	17.78	63	Favorite Indonesian dishes	✓	27.78	56	Evidence-Based Case Report examination for medical students	✗
	17.78	48	Thoughts on Academy of Champions (CoC for high schoolers)	✓	27.78	71	Research experiences in Kalimantan	✓

Table 9: Qualitative analysis of the Top-10 and Bottom-10 dialogues in the ID-EN and JV-ID-EN configurations based on the average performance of all evaluated LLMs. **Avg.** denotes the average model performance score, **ID** refers to the dialogue identifier, **Topic** describes the dialogue content, and **Too Specific (?)** indicates whether annotators considered the dialogue excessively specific due to local knowledge, personal experiences, or temporally grounded contexts.

This is a practical limitation worth noting when deploying smaller-context models in few-shot summarization pipelines.

G.2 Script Inconsistency in Algerian Arabic

For the AR-DZ-FR language pair, the expected output is *Arabizi* - a transliteration of Algerian Arabic using Latin characters mixed with numerals. However, we observe that several models occasionally produce summaries in Arabic script instead, despite the input dialogue being written in Arabizi. This script mismatch directly penalises surface-level metrics such as ROUGE, since character-level overlap between Arabic script and Arabizi is effectively zero even when the semantic content is correct. This explains the notably higher BERTScore-F values relative to ROUGE on this language pair - the model captures meaning but renders it in the wrong script. More broadly, Arabizi lacks a standardised orthography and is not formally supported in most model training pipelines, making consistent script adherence difficult to enforce without explicit instruction.

H Quantitative Results

This section provides a comprehensive overview of the quantitative findings obtained under all experimental configurations.

H.1 Question Answering

For the Question Answering task, results are analyzed from two complementary perspectives: (1) the effect of reasoning versus non-reasoning settings, and (2) the impact of zero-shot versus one-shot prompting strategies.

H.1.1 Reasoning vs. Non-Reasoning

Table 10 summarizes the performance comparison between reasoning-enabled and non-reasoning configurations.

H.1.2 0-shot vs. 1-shot

The comparative results for zero-shot and one-shot prompting setups are presented in Table 11.

H.2 Text Summarization

For the Text Summarization task, we similarly examine two major dimensions: the role of reasoning mechanisms and the effect of prompt shot size.

H.2.1 Reasoning vs. Non-Reasoning

Detailed results comparing reasoning and non-reasoning settings across different language combinations are reported in Table 12, Table 13, Table 14, Table 15, and Table 16.

H.2.2 0-shot vs. Few-shot

The overall results for the 0-shot versus few-shot setting are reported in Table 17, Table 18, Table 19,

	ID-EN		JV-ID-EN		SU-ID-EN		HA-EN		AR_DZ-FR	
model	**Reasoning**	**Non-reasoning**	**Reasoning**	**Non-reasoning**	**Reasoning**	**Non-reasoning**	**Reasoning**	**Non-reasoning**	**Reasoning**	**Non-reasoning**
Global										
Qwen2.5-3B-Instruct	11.51	5.85	25.85	18.98	86.66	76.16	43.93	43.94	91.91	25.25
Qwen2.5-7B-Instruct	12.12	4.64	35.75	24.64	90.50	86.68	32.82	43.93	40.40	38.37
Qwen3-4B	15.55	8.08	42.22	30.90	90.70	89.29	49.49	55.55	39.39	32.32
Qwen3-8B	21.41	8.68	46.66	31.71	89.89	89.29	50.00	45.95	41.41	44.44
Aya23-8B	25.65	8.28	31.51	23.03	86.46	80.00	29.79	28.78	28.28	14.14
Gemma2-9B-Instruct	13.73	11.91	45.05	43.63	91.71	91.91	82.82	82.82	52.52	50.50
Gemma3-4B-Instruct	25.05	23.43	41.01	36.76	91.31	90.70	71.21	69.19	36.36	39.39
Regional										
Sailor2-8B	35.55	27.67	37.17	40.40	86.66	86.46	-	-	-	-
Sahabat-AI-Gemma	16.16	13.73	46.06	43.23	93.13	93.53	-	-	-	-
SILMA-9B-Instruct	-	-	-	-	-	-	-	-	0.0	0.0
ALLAM-7B-Instruct	-	-	-	-	-	-	-	-	39.39	32.32

Table 10: Statistics (Acc. %) per-language combination of Reasoning vs. non-Reasoning approach on Question Answering task

	ID-EN		JV-ID-EN		SU-ID-EN		HA-EN		AR_DZ-FR	
model	**0-shot**	**1-shot**	**0-shot**	**1-shot**	**0-shot**	**1-shot**	**0-shot**	**1-shot**	**0-shot**	**1-shot**
Global										
Qwen2.5-3B-Instruct	5.85	4,64	18.08	19.19	76.16	72.32	43.93	42.42	25.25	25.25
Qwen2.5-7B-Instruct	4.64	3.63	24.64	24.24	88.68	87.07	43.93	42.92	38.38	33.33
Qwen3-4B	8.88	8.08	30.90	28.08	90.70	89.09	49.49	51.01	39.39	38.38
Qwen3-8B	5.65	7.07	31.71	31.31	89.89	89.29	50.00	44.44	41.41	37.37
Aya23-8B	8.28	13.93	23.03	28.08	80.0	84.24	28.78	26.76	14.14	13.13
Gemma2-9B-Instruct	11.91	9.09	43.63	39.19	91.91	91.51	82.82	81.81	50.50	49.49
Gemma3-4B-Instruct	23.43	20.40	36.76	33.93	90.70	90.50	69.19	62.62	39.39	38.38
Regional										
Sailor2-8B	27.67	29.89	40.40	39.19	86.46	85.45	-	-	-	-
Sahabat-AI-Gemma	13.73	12.12	43.23	41.01	93.53	93.53	-	-	-	-
SILMA-9B-Instruct	-	-	-	-	-	-	-	-	0.0	0.0
ALLAM-7B-Instruct	-	-	-	-	-	-	-	-	32.32	35.35

Table 11: Statistics (Acc. %) per-language combination of 0-shot vs. 1-shot approach on Question Answering task

Table 20, Table 21,

H.3 Human-written vs. Machine-generated

Table 22 reports the overall results comparing human-written vs. machine generated conversation.

I Additional Results

I.1 Taxonomy of Unanswerable QA Category

For the Question Answering (QA) task, annotators for each language were instructed to create up to five unanswerable questions per instance. This design ensures that each instance is represented across the five targeted categories: Negation, Antonym, Entity-Swap, Mutual-Exclusion, and Impossible-Condition. By the end of the annotation process, three language combinations (ID–EN, JV–ID–EN, and SU–ID–EN) successfully produced up to five unanswerable questions for each instance. Table 23 reports the model performance (Acc.%) on these three language combinations across all unanswerable categories. The results indicate that *Impossible-Condition* constitutes the most challenging category, whereas *Negation*, *Antonym*, and *Entity-Swap* are relatively easier for the models to handle.

		ID-EN								
model	**Reasoning**	**ROUGE1**	**ROUGE2**	**ROUGE3**	**ROUGE4**	**METEOR**	**CHRF++**	**BERTScore-p**	**BERTScore-r**	**BERTScore-f**
Global										
Qwen2.5-3B-Instruct	✓	0.292	0.078	0.027	0.010	0.216	32.022	0.903	0.896	0.895
Qwen2.5-3B-Instruct	✗	0.343	0.084	0.029	0.010	0.263	36.575	0.896	0.902	0.894
Qwen2.5-7B-Instruct	✓	0.337	0.102	0.042	0.017	0.227	32.942	0.910	0.899	0.899
Qwen2.5-7B-Instruct	✗	0.328	0.090	0.033	0.013	0.274	36.315	0.878	0.905	0.886
Qwen3-4B	✓	0.352	0.102	0.039	0.016	0.258	36.069	0.904	0.904	0.899
Qwen3-4B	✗	0.361	0.102	0.038	0.014	0.276	37.676	0.898	0.905	0.897
Qwen3-8B	✓	0.342	0.099	0.037	0.015	0.248	35.331	0.904	0.902	0.898
Qwen3-8B	✗	0.242	0.070	0.027	0.011	0.206	26.973	0.867	0.903	0.880
Aya23-8B	✓	0.203	0.053	0.020	0.008	0.170	27.000	0.894	0.890	0.888
Aya23-8B	✗	0.347	0.098	0.037	0.015	0.275	38.158	0.898	0.904	0.896
Gemma2-9B-Instruct	✓	0.348	0.114	0.049	0.024	0.253	35.622	0.909	0.905	0.902
Gemma2-9B-Instruct	✗	0.396	0.126	0.054	0.025	0.303	40.285	0.905	0.910	0.903
Gemma3-4B-Instruct	✓	0.365	0.106	0.040	0.017	0.285	38.663	0.897	0.904	0.897
Gemma3-4B-Instruct	✗	0.287	0.079	0.030	0.012	0.270	36.126	0.847	0.905	0.872
Regional										
Sailor2-8B	✓	0.243	0.059	0.020	0.007	0.169	25.686	0.749	0.750	0.745
Sailor2-8B	✗	0.278	0.067	0.023	0.007	0.203	29.451	0.732	0.746	0.736
Sahabat-AI-Gemma	✓	0.349	0.110	0.045	0.022	0.255	35.563	0.909	0.903	0.901
Sahabat-AI-Gemma	✗	0.370	0.122	0.048	0.021	0.293	39.021	0.901	0.907	0.899

Table 12: Statistics for the remaining evaluation metrics (ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, METEOR, CHRF++, and BERTScore-P/R/F) on the Indonesian-English (ID-EN) language pair, comparing reasoning and non-reasoning approaches for the Text Summarization task.

		JV-ID-EN								
model	**Reasoning**	**ROUGE1**	**ROUGE2**	**ROUGE3**	**ROUGE4**	**METEOR**	**CHRF++**	**BERTScore-p**	**BERTScore-r**	**BERTScore-f**
Global										
Qwen2.5-3B-Instruct	✓	0.298	0.071	0.023	0.008	0.189	30.166	0.904	0.884	0.892
Qwen2.5-3B-Instruct	✗	0.345	0.084	0.030	0.012	0.251	37.283	0.895	0.894	0.893
Qwen2.5-7B-Instruct	✓	0.312	0.088	0.032	0.011	0.187	29.252	0.910	0.886	0.896
Qwen2.5-7B-Instruct	✗	0.311	0.079	0.029	0.010	0.247	36.214	0.882	0.892	0.885
Qwen3-4B	✓	0.336	0.082	0.030	0.010	0.221	34.275	0.903	0.890	0.895
Qwen3-4B	✗	0.353	0.089	0.033	0.013	0.249	37.348	0.900	0.895	0.896
Qwen3-8B	✓	0.327	0.084	0.027	0.007	0.216	33.971	0.902	0.891	0.895
Qwen3-8B	✗	0.266	0.071	0.025	0.009	0.199	28.539	0.872	0.894	0.881
Aya23-8B	✓	0.200	0.047	0.016	0.005	0.143	24.163	0.891	0.878	0.883
Aya23-8B	✗	0.349	0.092	0.034	0.012	0.252	38.227	0.897	0.892	0.894
Gemma2-9B-Instruct	✓	0.319	0.086	0.030	0.011	0.206	31.614	0.906	0.889	0.896
Gemma2-9B-Instruct	✗	0.369	0.108	0.043	0.017	0.270	38.801	0.901	0.898	0.898
Gemma3-4B-Instruct	✓	0.349	0.085	0.029	0.012	0.248	37.752	0.895	0.894	0.893
Gemma3-4B-Instruct	✗	0.238	0.059	0.020	0.007	0.228	32.902	0.826	0.895	0.857
Regional										
Sailor2-8B	✓	0.207	0.039	0.011	0.002	0.128	22.042	0.700	0.692	0.695
Sailor2-8B	✗	0.225	0.047	0.015	0.004	0.156	25.473	0.658	0.664	0.660
Sahabat-AI-Gemma	✓	0.330	0.098	0.041	0.018	0.213	31.905	0.907	0.887	0.895
Sahabat-AI-Gemma	✗	0.354	0.107	0.044	0.018	0.256	37.021	0.900	0.893	0.895

Table 13: Statistics for the remaining evaluation metrics (ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, METEOR, CHRF++, and BERTScore-P/R/F) on the Javanese-Indonesian-English (JV-ID-EN) language pair, comparing reasoning and non-reasoning approaches for the Text Summarization task.

Prompt example for Question Answering

Answer the question provided according to the context in the dialog. Just write the answer in the form of A, B, C, D, or E based on the appropriate option, without including the sentences accompanying the options. If the question does not match the context of the conversation or if none of the answer choices are correct, choose option "<There is no correct answer>".

Below are <NUMBER OF SHOT> example QA pairs from the example dialog.

EXAMPLE DIALOG
<DIALOG EXAMPLE>

EXAMPLE <i>
QUESTION: <QUESTION EXAMPLE i>
OPTIONS:
A. <OPTION A>
B. <OPTION B>
C. <OPTION C>
D. <OPTION D>
E. <OPTION E>
ANSWER: <ANSWER EXAMPLE i>

Now answer the question for the dialog below.

DIALOG
<DIALOG>

QUESTION
<QUESTION>

OPTIONS
A. <OPTION A>
B. <OPTION B>
C. <OPTION C>
D. <OPTION D>
E. <OPTION E>

OUTPUT FORMAT
Return a JSON response in the following format:
<OUTPUT FORMAT>

OUTPUT

Figure 17: Prompt template for the Question Answering task (Multiple Choice subset). The highlighted middle section corresponds to an optional few-shot component and is included only in the few-shot prompting setting.

Prompt example for Dialogue Summarization

INSTRUCTIONS
Your primary task is to write a concise summary of a given dialogue. The summary should be focused on the provided topic and adhere strictly to the rules and output format specified below:
- The summary should be written in target-lang. Once again, write the summary in target-lang.
- The summary must be a single paragraph.
- The paragraph must be 3 to 5 sentences.
- The summary must be strictly factually correct, drawing all information directly from the dialogue. Do not introduce any external information or make assumptions not supported by the text.
- Prior to composing the summary, provide a brief analysis of the dialogue. This analysis should identify the key points selected for the summary and offer a justification for their inclusion.

EXAMPLES
DIALOGUE EXAMPLE
Topic: <TOPIC EXAMPLE>
Dialogue:
<DIALOGUE EXAMPLE>

SUMMARY EXAMPLE i
{
"summary": SUMMARY i
}

Now, please summarize the dialogue below.

DIALOGUE
Topic: <TOPIC>
Dialogue:
<DIALOGUE>

OUTPUT FORMAT
Return a JSON response in the following format:
<OUTPUT FORMAT>

OUTPUT

Figure 18: Prompt template for the Dialogue Summarization task. The highlighted middle section corresponds to an optional few-shot component and is included only in the few-shot prompting setting.

Prompt example for Topic Classification

INSTRUCTIONS
Your primary task is to classify the provided code-switched/code-mixed dialogue into one of the following categories:
- Entertainment
- Science/Technology
- Social/Culture
- Education
- Daily Life

Before classifying, please provide a brief analysis of the dialogue. This analysis should identify the key points that support your classification.

DIALOGUE-TOPIC PAIR EXAMPLE
Dialogue:
<DIALOG EXAMPLE>

{
"category": <CATEGORY EXAMPLE>
}

Now, please classify the dialogue below.

DIALOGUE
Dialogue:
<DIALOG>

OUTPUT FORMAT
Return a JSON response in the following format:
{
"explanation": "This section contains a brief analysis of the dialogue that supports the classification.",
"category": "This section contains exactly one category selected from the possible options. Output only the category name, without any additional text."
}

OUTPUT

Figure 19: Prompt template for the Topic Classification.

Prompt example for Machine Generated–Conversational Text

Write a dialogue about <TOPIC> that uses code-switching and fulfills the following conditions:
- The dialogue involves <NUMBER OF PERSON> people.
- Use <SPEAKER'S LABEL> to label the speakers.
- The language used in the conversation should be code-switched between <LANGUAGE>. Make sure that each language is present in the conversation.
- The conversation is informal, so the use of slang, casual expressions, and relaxed punctuation, grammar, or capitalization is allowed.
- The dialogue should consist of 50 to 150 utterances (about 15 minutes conversation).
- Emojis may be used in the dialogue.
- The dialogue takes place in an online setting (e.g., group chat, direct messages).
- Replies to specific messages are allowed and can be included.
- Ensure the conversation flows naturally.

Use the following format for the dialogue:
#1 A: insert text
#2 B: insert text
#3 C [reply to #1]: insert text
#4 A: insert text
#5 D: insert text
#6 A [reply to #5]: insert text

Write the dialogue here:

Figure 20: Prompt template for the Machine Generated–Conversational Text.

		SU-ID-EN								
model	**Reasoning**	**ROUGE1**	**ROUGE2**	**ROUGE3**	**ROUGE4**	**METEOR**	**CHRF++**	**BERTScore-p**	**BERTScore-r**	**BERTScore-f**
Global										
Qwen2.5-3B-Instruct	✓	0.346	0.109	0.043	0.017	0.219	32.329	0.914	0.890	0.899
Qwen2.5-3B-Instruct	✗	0.394	0.119	0.046	0.018	0.278	39.479	0.908	0.898	0.901
Qwen2.5-7B-Instruct	✓	0.348	0.119	0.051	0.024	0.203	30.222	0.920	0.889	0.902
Qwen2.5-7B-Instruct	✗	0.372	0.121	0.047	0.019	0.276	38.705	0.897	0.898	0.895
Qwen3-4B	✓	0.397	0.126	0.050	0.022	0.253	36.549	0.918	0.897	0.905
Qwen3-4B	✗	0.420	0.137	0.057	0.026	0.280	39.511	0.915	0.900	0.905
Qwen3-8B	✓	0.382	0.123	0.049	0.022	0.238	34.731	0.917	0.895	0.903
Qwen3-8B	✗	0.324	0.107	0.043	0.018	0.234	31.775	0.893	0.898	0.893
Aya23-8B	✓	0.264	0.082	0.035	0.015	0.171	27.423	0.907	0.883	0.892
Aya23-8B	✗	0.404	0.134	0.055	0.023	0.281	40.073	0.911	0.898	0.902
Gemma2-9B-Instruct	✓	0.391	0.141	0.063	0.028	0.252	34.637	0.922	0.895	0.906
Gemma2-9B-Instruct	✗	0.441	0.161	0.074	0.037	0.309	41.759	0.917	0.904	0.908
Gemma3-4B-Instruct	✓	0.422	0.144	0.059	0.028	0.299	40.948	0.914	0.903	0.906
Gemma3-4B-Instruct	✗	0.328	0.110	0.046	0.021	0.294	38.784	0.852	0.905	0.875
Regional										
Sailor2-8B	✓	0.210	0.062	0.021	0.008	0.128	19.015	0.592	0.581	0.585
Sailor2-8B	✗	0.258	0.075	0.026	0.010	0.172	24.393	0.588	0.590	0.588
Sahabat-AI-Gemma	✓	0.400	0.153	0.069	0.034	0.257	35.298	0.919	0.896	0.905
Sahabat-AI-Gemma	✗	0.448	0.163	0.073	0.037	0.309	41.698	0.917	0.902	0.907

Table 14: Statistics for the remaining evaluation metrics (ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, METEOR, CHRF++, and BERTScore-P/R/F) on the Sundanese-Indonesian-English (SU-ID-EN) language pair, comparing reasoning and non-reasoning approaches for the Text Summarization task.

		HA-EN								
model	**Reasoning**	**ROUGE1**	**ROUGE2**	**ROUGE3**	**ROUGE4**	**METEOR**	**CHRF++**	**BERTScore-p**	**BERTScore-r**	**BERTScore-f**
Global										
Qwen2.5-3B-Instruct	✓	0.060	0.007	0.002	0.000	0.059	9.222	0.757	0.788	0.770
Qwen2.5-3B-Instruct	✗	0.063	0.009	0.002	0.001	0.071	9.568	0.758	0.802	0.777
Qwen2.5-7B-Instruct	✓	0.216	0.032	0.006	0.001	0.130	19.915	0.842	0.836	0.837
Qwen2.5-7B-Instruct	✗	0.176	0.029	0.009	0.003	0.131	18.438	0.814	0.835	0.822
Qwen3-4B	✓	0.062	0.007	0.001	0.000	0.047	7.423	0.758	0.776	0.765
Qwen3-4B	✗	0.077	0.011	0.003	0.002	0.066	9.369	0.764	0.789	0.774
Qwen3-8B	✓	0.038	0.005	0.001	0.000	0.031	4.755	0.756	0.764	0.758
Qwen3-8B	✗	0.017	0.004	0.001	0.000	0.024	3.206	0.752	0.772	0.759
Aya23-8B	✓	0.127	0.021	0.006	0.001	0.111	17.010	0.817	0.824	0.818
Aya23-8B	✗	0.097	0.017	0.005	0.002	0.097	14.827	0.792	0.820	0.803
Gemma2-9B-Instruct	✓	0.350	0.094	0.033	0.012	0.212	30.277	0.891	0.871	0.879
Gemma2-9B-Instruct	✗	0.360	0.098	0.036	0.015	0.243	33.598	0.884	0.878	0.879
Gemma3-4B-Instruct	✓	0.261	0.049	0.013	0.003	0.188	26.318	0.841	0.859	0.848
Gemma3-4B-Instruct	✗	0.160	0.032	0.009	0.003	0.150	19.467	0.805	0.856	0.827

Table 15: Statistics for the remaining evaluation metrics (ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, METEOR, CHRF++, and BERTScore-P/R/F) on the Hausa-English (HA-EN) language pair, comparing reasoning and non-reasoning approaches for the Text Summarization task.

		AR_DZ-FR								
model	**Reasoning**	**ROUGE1**	**ROUGE2**	**ROUGE3**	**ROUGE4**	**METEOR**	**CHRF++**	**BERTScore-p**	**BERTScore-r**	**BERTScore-f**
Global										
Qwen2.5-3B-Instruct	✓	0.018	0.001	0.000	0.000	0.039	3.756	0.806	0.781	0.792
Qwen2.5-3B-Instruct	✗	0.006	0.001	0.001	0.001	0.032	1.184	0.807	0.766	0.784
Qwen2.5-7B-Instruct	✓	0.149	0.022	0.006	0.002	0.109	19.202	0.850	0.845	0.847
Qwen2.5-7B-Instruct	✗	0.083	0.015	0.005	0.002	0.091	11.835	0.797	0.830	0.812
Qwen3-4B	✓	0.093	0.015	0.004	0.001	0.076	12.491	0.815	0.813	0.812
Qwen3-4B	✗	0.087	0.019	0.006	0.002	0.078	11.808	0.794	0.814	0.802
Qwen3-8B	✓	0.086	0.013	0.002	0.000	0.076	12.815	0.802	0.819	0.809
Qwen3-8B	✗	0.025	0.005	0.002	0.000	0.033	4.445	0.773	0.805	0.787
Aya23-8B	✓	0.054	0.009	0.002	0.001	0.073	14.529	0.820	0.783	0.800
Aya23-8B	✗	0.042	0.010	0.003	0.001	0.068	9.450	0.798	0.779	0.786
Gemma2-9B-Instruct	✓	0.138	0.027	0.009	0.003	0.100	16.490	0.838	0.822	0.829
Gemma2-9B-Instruct	✗	0.108	0.019	0.007	0.002	0.107	16.018	0.810	0.817	0.812
Gemma3-4B-Instruct	✓	0.033	0.005	0.001	0.000	0.048	4.412	0.787	0.787	0.785
Gemma3-4B-Instruct	✗	0.020	0.004	0.002	0.001	0.032	2.463	0.758	0.776	0.764
Regional										
SILMA-9B-Instruct	✓	0.025	0.007	0.003	0.001	0.030	4.326	0.798	0.765	0.779
SILMA-9B-Instruct	✗	0.010	0.002	0.000	0.000	0.010	1.230	0.715	0.756	0.733
ALLAM-7B-Instruct	✓	0.007	0.002	0.001	0.000	0.018	1.141	0.381	0.365	0.372
ALLAM-7B-Instruct	✗	0.004	0.000	0.000	0.000	0.014	0.460	0.361	0.358	0.359

Table 16: Statistics for the remaining evaluation metrics (ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, METEOR, CHRF++, and BERTScore-P/R/F) on the Algerian Arabic-France (ar_DZ-FR) language pair, comparing reasoning and non-reasoning approaches for the Text Summarization task.

		ID-EN								
model	**N-shot**	**ROUGE1**	**ROUGE2**	**ROUGE3**	**ROUGE4**	**METEOR**	**CHRF++**	**BERTScore-p**	**BERTScore-r**	**BERTScore-f**
Global										
Qwen2.5-3B-Instruct	0	0.343	0.084	0.029	0.010	0.263	36.575	0.896	0.902	0.894
Qwen2.5-3B-Instruct	1	0.364	0.102	0.038	0.014	0.270	36.502	0.903	0.904	0.899
Qwen2.5-3B-Instruct	3	0.364	0.102	0.040	0.014	0.273	37.287	0.903	0.905	0.900
Qwen2.5-7B-Instruct	0	0.328	0.090	0.033	0.013	0.274	36.315	0.878	0.905	0.886
Qwen2.5-7B-Instruct	1	0.368	0.117	0.047	0.017	0.272	36.228	0.906	0.905	0.901
Qwen2.5-7B-Instruct	3	0.377	0.120	0.051	0.021	0.275	36.748	0.909	0.906	0.902
Qwen3-4B	0	0.303	0.087	0.034	0.015	0.254	34.148	0.881	0.903	0.887
Qwen3-4B	1	0.301	0.093	0.038	0.015	0.219	30.319	0.891	0.883	0.883
Qwen3-4B	3	0.177	0.057	0.024	0.009	0.136	18.194	0.854	0.842	0.845
Qwen3-8B	0	0.250	0.073	0.028	0.011	0.229	29.557	0.866	0.902	0.879
Qwen3-8B	1	0.366	0.114	0.044	0.018	0.267	37.242	0.907	0.906	0.901
Qwen3-8B	3	0.359	0.112	0.046	0.018	0.264	36.605	0.906	0.905	0.900
Aya23-8B	0	0.347	0.098	0.037	0.015	0.275	38.158	0.898	0.904	0.896
Aya23-8B	1	0.370	0.112	0.043	0.018	0.276	38.069	0.903	0.905	0.900
Aya23-8B	3	0.374	0.111	0.047	0.018	0.280	38.010	0.904	0.906	0.900
Gemma2-9B-Instruct	0	0.396	0.126	0.054	0.025	0.303	40.285	0.905	0.910	0.903
Gemma2-9B-Instruct	1	0.408	0.139	0.061	0.028	0.310	40.324	0.910	0.911	0.906
Gemma2-9B-Instruct	3	0.406	0.131	0.058	0.026	0.302	39.897	0.910	0.910	0.906
Gemma3-4B-Instruct	0	0.287	0.079	0.030	0.012	0.270	36.126	0.847	0.905	0.872
Gemma3-4B-Instruct	1	0.348	0.101	0.039	0.016	0.298	40.053	0.871	0.906	0.885
Gemma3-4B-Instruct	3	0.357	0.105	0.039	0.015	0.298	40.241	0.873	0.906	0.886
Regional										
Sailor2-8B	0	0.278	0.067	0.023	0.007	0.203	29.451	0.732	0.746	0.736
Sailor2-8B	1	0.240	0.063	0.023	0.008	0.180	25.247	0.636	0.646	0.639
Sailor2-8B	3	0.159	0.043	0.015	0.004	0.119	16.885	0.412	0.419	0.414
Sahabat-AI-Gemma	0	0.370	0.122	0.048	0.021	0.293	39.021	0.901	0.907	0.899
Sahabat-AI-Gemma	1	0.389	0.134	0.061	0.027	0.292	37.878	0.912	0.908	0.906
Sahabat-AI-Gemma	3	0.385	0.128	0.056	0.025	0.286	38.475	0.911	0.907	0.904

Table 17: Statistics for the remaining evaluation metrics (ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, METEOR, CHRF++, and BERTScore-P/R/F) on the Indonesian-English (ID-EN) language pair, comparing 0-shot and few-shot approaches for the Text Summarization task.

		JV-ID-EN								
model	**N-shot**	**ROUGE1**	**ROUGE2**	**ROUGE3**	**ROUGE4**	**METEOR**	**CHRF++**	**BERTScore-p**	**BERTScore-r**	**BERTScore-f**
Global										
Qwen2.5-3B-Instruct	0	0.345	0.084	0.030	0.012	0.251	37.283	0.895	0.894	0.893
Qwen2.5-3B-Instruct	1	0.371	0.117	0.060	0.036	0.263	38.067	0.903	0.898	0.899
Qwen2.5-3B-Instruct	3	0.372	0.121	0.060	0.032	0.273	38.564	0.904	0.899	0.901
Qwen2.5-7B-Instruct	0	0.311	0.079	0.029	0.010	0.247	36.214	0.882	0.892	0.885
Qwen2.5-7B-Instruct	1	0.361	0.108	0.047	0.022	0.243	36.253	0.905	0.896	0.899
Qwen2.5-7B-Instruct	3	0.374	0.122	0.058	0.029	0.255	36.736	0.909	0.899	0.903
Qwen3-4B	0	0.321	0.086	0.034	0.014	0.251	35.304	0.885	0.895	0.888
Qwen3-4B	1	0.329	0.113	0.061	0.038	0.240	34.266	0.890	0.881	0.885
Qwen3-4B	3	0.392	0.129	0.062	0.035	0.287	40.340	0.907	0.903	0.904
Qwen3-8B	0	0.238	0.060	0.022	0.008	0.202	28.125	0.860	0.891	0.873
Qwen3-8B	1	0.357	0.097	0.038	0.016	0.241	36.627	0.904	0.896	0.898
Qwen3-8B	3	0.381	0.131	0.061	0.030	0.257	36.889	0.910	0.898	0.903
Aya23-8B	0	0.349	0.092	0.034	0.012	0.252	38.227	0.897	0.892	0.894
Aya23-8B	1	0.369	0.119	0.055	0.030	0.273	39.667	0.900	0.898	0.898
Aya23-8B	3	0.376	0.123	0.057	0.031	0.283	40.235	0.901	0.900	0.899
Gemma2-9B-Instruct	0	0.369	0.108	0.043	0.017	0.270	38.801	0.901	0.898	0.898
Gemma2-9B-Instruct	1	0.387	0.119	0.047	0.019	0.276	39.167	0.906	0.900	0.902
Gemma2-9B-Instruct	3	0.389	0.123	0.050	0.019	0.279	39.567	0.908	0.901	0.903
Gemma3-4B-Instruct	0	0.238	0.059	0.020	0.007	0.228	32.902	0.826	0.895	0.857
Gemma3-4B-Instruct	1	0.330	0.086	0.030	0.011	0.278	38.882	0.867	0.900	0.881
Gemma3-4B-Instruct	3	0.331	0.095	0.039	0.017	0.279	39.480	0.867	0.900	0.882
Regional										
Sailor2-8B	0	0.225	0.047	0.015	0.004	0.156	25.473	0.658	0.664	0.660
Sailor2-8B	1	0.203	0.052	0.019	0.006	0.139	22.138	0.573	0.576	0.574
Sailor2-8B	3	0.135	0.034	0.013	0.005	0.091	14.443	0.368	0.370	0.369
Sahabat-AI-Gemma	0	0.354	0.107	0.044	0.018	0.256	37.021	0.900	0.893	0.895
Sahabat-AI-Gemma	1	0.382	0.133	0.064	0.035	0.267	36.991	0.910	0.897	0.902
Sahabat-AI-Gemma	3	0.398	0.148	0.077	0.042	0.279	38.124	0.912	0.899	0.904

Table 18: Statistics for the remaining evaluation metrics (ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, METEOR, CHRF++, and BERTScore-P/R/F) on the Javanese-Indonesian-English (JV-ID-EN) language pair, comparing 0-shot and few-shot approaches for the Text Summarization task.

		SU-ID-EN								
model	**N-shot**	**ROUGE1**	**ROUGE2**	**ROUGE3**	**ROUGE4**	**METEOR**	**CHRF++**	**BERTScore-p**	**BERTScore-r**	**BERTScore-f**
Global										
Qwen2.5-3B-Instruct	0	0.394	0.119	0.046	0.018	0.278	39.479	0.908	0.898	0.901
Qwen2.5-3B-Instruct	1	0.372	0.114	0.044	0.018	0.257	36.972	0.910	0.896	0.901
Qwen2.5-3B-Instruct	3	0.399	0.123	0.051	0.022	0.275	38.956	0.910	0.898	0.902
Qwen2.5-7B-Instruct	0	0.372	0.121	0.047	0.019	0.276	38.705	0.897	0.898	0.895
Qwen2.5-7B-Instruct	1	0.382	0.130	0.055	0.025	0.252	36.277	0.916	0.896	0.904
Qwen2.5-7B-Instruct	3	0.408	0.138	0.057	0.025	0.273	38.435	0.916	0.899	0.906
Qwen3-4B	0	0.392	0.133	0.053	0.024	0.284	37.895	0.904	0.900	0.900
Qwen3-4B	1	0.270	0.093	0.043	0.020	0.192	26.580	0.881	0.854	0.866
Qwen3-4B	3	0.089	0.028	0.012	0.005	0.065	9.357	0.834	0.795	0.813
Qwen3-8B	0	0.331	0.110	0.047	0.023	0.256	33.668	0.892	0.898	0.892
Qwen3-8B	1	0.404	0.132	0.054	0.024	0.273	38.908	0.916	0.900	0.905
Qwen3-8B	3	0.420	0.143	0.060	0.028	0.285	39.631	0.919	0.902	0.908
Aya23-8B	0	0.404	0.134	0.055	0.023	0.281	40.073	0.911	0.898	0.902
Aya23-8B	1	0.375	0.114	0.044	0.019	0.269	39.839	0.905	0.896	0.898
Aya23-8B	3	0.385	0.124	0.051	0.022	0.278	40.240	0.906	0.897	0.900
Gemma2-9B-Instruct	0	0.441	0.161	0.074	0.037	0.309	41.759	0.917	0.904	0.908
Gemma2-9B-Instruct	1	0.431	0.156	0.071	0.036	0.306	41.826	0.919	0.904	0.909
Gemma2-9B-Instruct	3	0.453	0.170	0.080	0.040	0.319	42.777	0.920	0.906	0.910
Gemma3-4B-Instruct	0	0.328	0.110	0.046	0.021	0.294	38.784	0.852	0.905	0.875
Gemma3-4B-Instruct	1	0.387	0.128	0.053	0.026	0.319	42.294	0.879	0.906	0.889
Gemma3-4B-Instruct	3	0.380	0.124	0.051	0.024	0.308	41.509	0.873	0.904	0.886
Regional										
Sailor2-8B	0	0.258	0.075	0.026	0.010	0.172	24.393	0.588	0.590	0.588
Sailor2-8B	1	0.224	0.064	0.021	0.008	0.152	22.468	0.545	0.545	0.544
Sailor2-8B	3	0.099	0.031	0.013	0.006	0.070	9.829	0.237	0.237	0.236
Sahabat-AI-Gemma	0	0.448	0.163	0.073	0.037	0.309	41.698	0.917	0.902	0.907
Sahabat-AI-Gemma	1	0.435	0.171	0.083	0.046	0.304	40.211	0.923	0.903	0.910
Sahabat-AI-Gemma	3	0.450	0.177	0.084	0.042	0.312	41.001	0.923	0.903	0.910

Table 19: Statistics for the remaining evaluation metrics (ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, METEOR, CHRF++, and BERTScore-P/R/F) on the Sundanese-Indonesian-English (SU-ID-EN) language pair, comparing 0-shot and few-shot approaches for the Text Summarization task.

		HA-EN								
model	**N-shot**	**ROUGE1**	**ROUGE2**	**ROUGE3**	**ROUGE4**	**METEOR**	**CHRF++**	**BERTScore-p**	**BERTScore-r**	**BERTScore-f**
Global										
Qwen2.5-3B-Instruct	0	0.063	0.009	0.002	0.001	0.071	9.568	0.758	0.802	0.777
Qwen2.5-3B-Instruct	1	0.207	0.041	0.015	0.005	0.151	21.180	0.829	0.848	0.837
Qwen2.5-3B-Instruct	3	0.269	0.062	0.027	0.014	0.185	26.410	0.845	0.856	0.849
Qwen2.5-7B-Instruct	0	0.176	0.029	0.009	0.003	0.131	18.438	0.814	0.835	0.822
Qwen2.5-7B-Instruct	1	0.284	0.064	0.025	0.010	0.170	23.237	0.860	0.853	0.855
Qwen2.5-7B-Instruct	3	0.275	0.063	0.027	0.012	0.172	24.478	0.852	0.852	0.851
Qwen3-4B	0	0.033	0.003	0.001	0.000	0.040	6.047	0.740	0.761	0.748
Qwen3-4B	1	0.137	0.026	0.008	0.003	0.111	14.864	0.788	0.805	0.794
Qwen3-4B	3	0.139	0.038	0.018	0.010	0.110	15.138	0.785	0.802	0.791
Qwen3-8B	0	0.020	0.002	0.000	0.000	0.036	5.450	0.751	0.762	0.754
Qwen3-8B	1	0.060	0.013	0.005	0.002	0.075	8.675	0.756	0.788	0.769
Qwen3-8B	3	0.163	0.034	0.017	0.008	0.124	17.889	0.798	0.824	0.809
Aya23-8B	0	0.097	0.017	0.005	0.002	0.097	14.827	0.792	0.820	0.803
Aya23-8B	1	0.082	0.013	0.003	0.001	0.093	11.774	0.775	0.815	0.792
Aya23-8B	3	0.219	0.038	0.013	0.002	0.151	22.700	0.821	0.842	0.830
Gemma2-9B-Instruct	0	0.360	0.098	0.036	0.015	0.243	33.598	0.884	0.878	0.879
Gemma2-9B-Instruct	1	0.397	0.126	0.053	0.024	0.273	35.381	0.893	0.885	0.888
Gemma2-9B-Instruct	3	0.397	0.128	0.055	0.026	0.272	35.009	0.895	0.885	0.888
Gemma3-4B-Instruct	0	0.160	0.032	0.009	0.003	0.150	19.467	0.805	0.856	0.827
Gemma3-4B-Instruct	1	0.196	0.048	0.017	0.006	0.175	21.231	0.806	0.862	0.830
Gemma3-4B-Instruct	3	0.220	0.053	0.019	0.007	0.185	23.459	0.816	0.865	0.837

Table 20: Statistics for the remaining evaluation metrics (ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, METEOR, CHRF++, and BERTScore-P/R/F) on the Hausa-English (HA-EN) language pair, comparing 0-shot and few-shot approaches for the Text Summarization task.

		AR_DZ-FR								
model	**N-shot**	**ROUGE1**	**ROUGE2**	**ROUGE3**	**ROUGE4**	**METEOR**	**CHRF++**	**BERTScore-p**	**BERTScore-r**	**BERTScore-f**
Global										
Qwen2.5-3B-Instruct	0	0.006	0.001	0.001	0.001	0.032	1.184	0.807	0.766	0.784
Qwen2.5-3B-Instruct	1	0.130	0.019	0.007	0.004	0.114	18.052	0.825	0.845	0.834
Qwen2.5-3B-Instruct	3	0.164	0.019	0.004	0.001	0.142	20.782	0.838	0.855	0.846
Qwen2.5-7B-Instruct	0	0.083	0.015	0.005	0.002	0.091	11.835	0.797	0.830	0.812
Qwen2.5-7B-Instruct	1	0.153	0.021	0.007	0.002	0.132	21.770	0.837	0.855	0.845
Qwen2.5-7B-Instruct	3	0.169	0.031	0.011	0.003	0.150	21.013	0.829	0.852	0.839
Qwen3-4B	0	0.069	0.015	0.005	0.002	0.083	10.413	0.784	0.823	0.802
Qwen3-4B	1	0.052	0.005	0.002	0.000	0.050	8.575	0.812	0.776	0.791
Qwen3-4B	3	0.029	0.003	0.001	0.000	0.032	5.063	0.810	0.749	0.776
Qwen3-8B	0	0.018	0.003	0.001	0.000	0.036	4.915	0.759	0.794	0.774
Qwen3-8B	1	0.015	0.002	0.001	0.000	0.025	4.285	0.799	0.757	0.774
Qwen3-8B	3	0.009	0.000	0.000	0.000	0.013	2.956	0.808	0.734	0.767
Aya23-8B	0	0.042	0.010	0.003	0.001	0.068	9.450	0.798	0.779	0.786
Aya23-8B	1	0.137	0.007	0.001	0.000	0.119	20.754	0.830	0.845	0.837
Aya23-8B	3	0.142	0.007	0.001	0.000	0.126	20.508	0.831	0.852	0.840
Gemma2-9B-Instruct	0	0.108	0.019	0.007	0.002	0.107	16.018	0.810	0.817	0.812
Gemma2-9B-Instruct	1	0.109	0.017	0.006	0.002	0.106	16.043	0.807	0.835	0.819
Gemma2-9B-Instruct	3	0.119	0.021	0.008	0.003	0.116	16.058	0.806	0.837	0.819
Gemma3-4B-Instruct	0	0.020	0.004	0.002	0.001	0.032	2.463	0.758	0.776	0.764
Gemma3-4B-Instruct	1	0.060	0.010	0.003	0.001	0.079	11.229	0.779	0.841	0.807
Gemma3-4B-Instruct	3	0.093	0.014	0.004	0.001	0.108	14.957	0.798	0.846	0.820
Regional										
SILMA-9B-Instruct	0	0.010	0.002	0.000	0.000	0.010	1.230	0.715	0.756	0.733
SILMA-9B-Instruct	1	0.041	0.007	0.002	0.000	0.027	4.038	0.740	0.779	0.756
SILMA-9B-Instruct	3	0.062	0.008	0.002	0.001	0.049	6.571	0.758	0.788	0.771
ALLAM-7B-Instruct	0	0.004	0.000	0.000	0.000	0.014	0.460	0.361	0.358	0.359
ALLAM-7B-Instruct	1	0.011	0.002	0.001	0.001	0.014	1.410	0.204	0.207	0.205
ALLAM-7B-Instruct	3	0.000	0.000	0.000	0.000	0.000	0.010	0.008	0.008	0.008

Table 21: Statistics for the remaining evaluation metrics (ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, METEOR, CHRF++, and BERTScore-P/R/F) on the Algerian Arabic-French (ar_DZ-FR) language pair, comparing 0-shot and few-shot approaches for the Text Summarization task.

	ID-EN		JV-ID-EN		SU-ID-EN		HA-EN		AR_DZ-FR	
Metric	**Human-written**	**Machine-gen**	**Human-written**	**Machine-gen**	**Human-written**	**Machine-gen**	**Human-written**	**Machine-gen**	**Human-written**	**Machine-gen**
2 speakers (50 dialogue)										
Avg length variance (tokens)	83.436	29.438	47.58	43.376	104.202	27.267	48.252	28.469	49.231	47151
Total replies	482	1	1049	9	831	1	747	2	182	0
Avg degree of reply distance	3.053	0.02	2.972	0.18	2.066	0.02	2.645	0.02	2.909	0
Avg imbalance ratio of speaker turns	1.364	1.016	1.315	1.019	1.368	1.017	1.195	1.017	1.587	1.026
Avg CMI	0.491	0.711	0.467	0.733	0.734	0.719	0.37	0.374	0.565	0.169
Avg SPF	0.306	0.423	0.312	0.441	0.461	0.436	0.202	0.189	0.251	0.075
Human preference	2.721	2.694	2.627	2.373	2.840	2.120	2.483	2.470	2.969	1.208
3 speakers (25 dialogue)										
Avg length variance (tokens)	44.122	22.732	50.32	42.88	98.882	23.811	48.519	23.069	54.157	38.657
Total replies	414	23	772	17	941	22	761	20	270	17
Avg degree of reply distance	4.309	0.98	3.814	0.558	3.222	0.668	3.188	0.723	3.921	0.76
Avg imbalance ratio of speaker turns	3.286	1.053	1.895	1.107	1.832	1.044	3.324	1.052	4.347	1.024
Avg CMI	0.444	0.704	0.476	0.689	0.77	0.718	0.348	0.202	0.565	0.117
Avg SPF	0.291	0.429	0.328	0.411	0.469	0.437	0.187	0.1	0.248	0.044
Human preference	2.692	2.538	2.603	2.359	2.893	2.107	2.467	2.520	2.942	1.192
4 speakers (25 dialogue)										
Avg length variance (tokens)	63.764	32.003	35.918	32.82	167.263	24.299	58.806	15.9	80.132	40.934
Total replies	718	43	950	39	1307	55	1062	43	305	45
Avg degree of reply distance	4.343	1.263	4.037	1.099	3.841	0.998	3.543	1.005	5.115	1.097
Avg imbalance ratio of speaker turns	3.459	1.1	2.586	1.209	2.3	1.098	2.786	1.086	4.272	1.097
Avg CMI	0.46	0.667	0.456	0.729	0.789	0.719	0.321	0.191	0.576	0.108
Avg SPF	0.298	0.393	0.318	0.44	0.479	0.452	0.172	0.096	0.261	0.041
Human preference	2.573	2.533	2.625	2.083	2.853	2.200	2.385	2.538	2.942	1.192

Table 22: Full quantitative statistics per-language combination of human written VS machine generated conversational pattern

	ID-EN					JV-ID-EN					SU-ID-EN				
Model	**Neg.**	**Ant.**	**Ent.**	**Mut.**	**Imp.**	**Neg.**	**Ant.**	**Ent.**	**Mut.**	**Imp.**	**Neg.**	**Ant.**	**Ent.**	**Mut.**	**Imp.**
Global															
Qwen2.5-3B-Instruct	46.46	45.45	36.36	59.59	76.76	58.58	46.46	42.42	67.67	88.88	90.90	77.77	77.77	88.88	95.95
Qwen2.5-7B-Instruct	48.48	61.61	40.40	63.63	87.87	60.60	51.51	52.52	65.65	90.90	80.80	68.68	76.76	78.78	91.91
Qwen3-4B	28.28	29.29	10.10	44.44	76.76	30.30	31.31	18.18	42.42	78.78	82.82	65.65	73.73	75.75	90.90
Qwen3-8B	33.83	44.44	23.23	73.73	88.88	50.50	41.41	33.33	68.68	90.90	93.93	86.86	86.86	85.85	98.98
Aya23-8B	54.54	44.44	38.38	64.64	77.77	50.50	43.43	38.38	65.65	81.81	72.72	66.66	52.52	73.73	79.79
Gemma2-9B-Instruct	35.35	31.31	20.20	75.75	87.87	45.45	36.36	20.20	74.74	90.90	92.92	86.86	87.87	92.92	94.94
Gemma3-4B-Instruct	10.10	6.06	4.04	18.18	31.31	13.13	10.10	7.07	23.23	34.34	51.51	26.26	42.42	40.40	54.54
Regional															
Sailor2-8B	10.10	12.12	3.03	15.15	16.16	13.13	7.07	3.03	16.16	12.12	41.41	26.26	23.23	28.28	34.34
Sahabat-AI-Gemma	40.40	25.25	19.19	74.74	86.86	44.44	39.39	18.18	77.77	91.91	92.92	85.85	82.82	91.91	92.92

Table 23: Statistics (Acc. %) per-language combination of Unanswerable question's category on Question Answering task. (Neg. = Negation, Ant. = Antonym, Ent. = Entity-Swap, Mut. = Mutually-Exclusion, and Imp. = Impossible-Condition)