NormGenesis: Multicultural Dialogue Generation via Exemplar-Guided Social Norm Modeling and Violation Recovery

Minki Hong[1] **and Jangho Choi**[1] **and Jihie Kim**[1]*
[1]Department of Computer Science and Artificial Intelligence, Dongguk University
{jackyh1, 2025120382}@dgu.ac.kr, jihie.kim@dgu.edu

Abstract

Social norms govern culturally appropriate behavior in communication, enabling dialogue systems to produce responses that are not only coherent but also socially acceptable. We present NormGenesis, a multicultural framework for generating and annotating socially grounded dialogues across English, Chinese, and Korean. To model the dynamics of social interaction beyond static norm classification, we propose a novel dialogue type, Violation-to-Resolution (V2R), which models the progression of conversations following norm violations through recognition and socially appropriate repair. To improve pragmatic consistency in underrepresented languages, we implement an exemplar-based iterative refinement early in the dialogue synthesis process. This design introduces alignment with linguistic, emotional, and sociocultural expectations before full dialogue generation begins. Using this framework, we construct a dataset of 10,800 multi-turn dialogues annotated at the turn level for norm adherence, speaker intent, and emotional response. Human and LLM-based evaluations demonstrate that NormGenesis significantly outperforms existing datasets in refinement quality, dialogue naturalness, and generalization performance. We show that models trained on our V2R-augmented data exhibit improved pragmatic competence in ethically sensitive contexts. Our work establishes a new benchmark for culturally adaptive dialogue modeling and provides a scalable methodology for norm-aware generation across linguistically and culturally diverse languages.

1 Introduction

Social norms are culturally defined expectations that guide appropriate behavior in specific contexts (Elster, 2006; Malle et al., 2014). In human communication, social norms support politeness, empathy, and social harmony. For dialogue systems, aligning with social norms enables responses that transcend syntactic correctness or task completion, contributing to pragmatic and interpersonal appropriateness (Kim et al., 2022). As conversational agents are increasingly deployed in socially embedded and open-domain settings, the ability to recognize and adhere to cultural norms has become a critical indicator of social and pragmatic competence (Zhan et al., 2023).

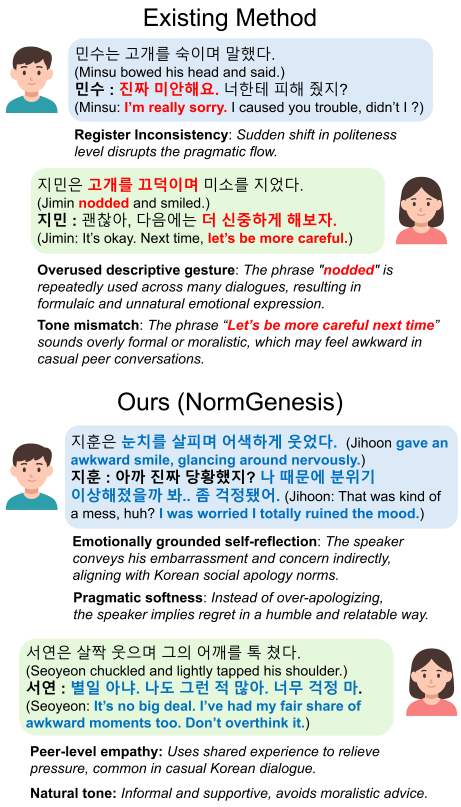

Figure 1: Comparison of generation outputs in Korean. Prior methods (Li et al., 2023) produce pragmatically inconsistent responses, including honorific misuse and unnatural tone (highlighted in red). In contrast, our framework yields culturally and pragmatically coherent outputs (highlighted in blue).

* Corresponding author: jihie.kim@dgu.edu
Code available at: https://github.com/bk123477/NormGenesis

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33793–33831
November 4-9, 2025 ©2025 Association for Computational Linguistics

Recent studies have integrated social norms into dialogue datasets and language models. Prior works have explored moral reasoning in language models (Forbes et al., 2020), norm-based labeling (Li et al., 2023), emotion-informed norm interpretation (Zhan et al., 2024), and cross-cultural generalization (Rao et al., 2025). While these efforts lay foundational groundwork for norm-aware generation, they primarily focus on English, a high-resource language. Although Chinese has received growing attention, the modeling of culturally appropriate behavior remains underdeveloped for low-resource languages. This limitation is especially pronounced in Korean, where existing models frequently exhibit inconsistencies in honorific usage, inadequate emotional alignment, and misrepresentation of role-based social dynamics (Jang et al., 2024; Lee et al., 2024), as illustrated in Figure 1.

To address the cultural and pragmatic limitations of existing dialogue datasets, we present a multicultural framework for generating and refining socially grounded dialogues across English, Chinese, and Korean. While English benefits from extensive data and modeling maturity, low-resource languages generation remains challenged by pragmatic mismatches, especially in tone and formality (Zhong et al., 2024). We mitigate this gap through an exemplar-based iterative refinement strategy. Given a target scenario, the system retrieves semantically and structurally aligned exemplars, using features such as intent, emotional tone, and discourse patterns (e.g., speaker roles and adjacency). These exemplars guide revision to ensure cultural alignment without requiring large-scale human annotation. We further introduce a novel dialogue category, *Violation-to-Resolution* (V2R), which captures how speakers recover from norm violations through contextually appropriate repair. This enables the modeling of pragmatically dynamic interactions that reflect both norm compliance and social recovery mechanisms. Leveraging our framework, we construct a dataset of 10,800 high-quality dialogues annotated at the turn level with norm adherence, violation, speaker intent, and emotional response, grounded in dialogue act theory (Bunt et al., 2020). We evaluate our approach through both human and LLM-based assessments of refinement quality, dialogue fluency, social appropriateness, and generalization. Experimental results show that models trained on our data significantly outperform existing baselines in socially complex and emotionally sensitive scenarios. These findings demonstrate the efficacy of our framework for enabling culturally adaptive dialogue generation across typologically diverse languages.

Our main contributions are as follows:

1. We present a multicultural framework for generating socially grounded dialogues in English, Chinese, and Korean. To address cultural and pragmatic degradation in low-resource settings, we propose an exemplar-based iterative refinement strategy using semantically relevant exemplars.

2. We propose a novel dialogue type, *Violation-to-Resolution* (V2R), that models how norm violations are followed by socially appropriate repair, enabling the representation of dynamic and culturally meaningful interaction patterns.

3. We construct a dataset of 10,800 multicultural dialogues with turn-level annotations for norm adherence, speaker intent, and emotional response, grounded in dialogue act theory.

4. We show that models trained on our dataset outperform prior resources in norm alignment, emotional coherence, and repair, as confirmed by both human and automatic evaluations.

2 Related Work

2.1 Social Norms in Dialogue Systems

Integrating social norms into dialogue systems is critical for generating contextually appropriate and socially coherent responses. Early work (Forbes et al., 2020) provided normative signals via moral judgments but lacked dialogue structure. Later efforts (Li et al., 2023) added norm annotations to multi-turn dialogues, focusing on adherence classification without modeling responses to violations. Recent work (Zhan et al., 2024) has begun modeling norm repair in dialogue. However, it remains monolingual and lacks fine-grained annotations capturing speaker intent and emotional response.

To bridge this gap, we introduce *Violation-to-Resolution* as a distinct response type, capturing repair strategies such as apology and explanation. Also, our framework supports this with turn-level annotations of communicative intent and emotional state, enabling more nuanced and socially competent generation.

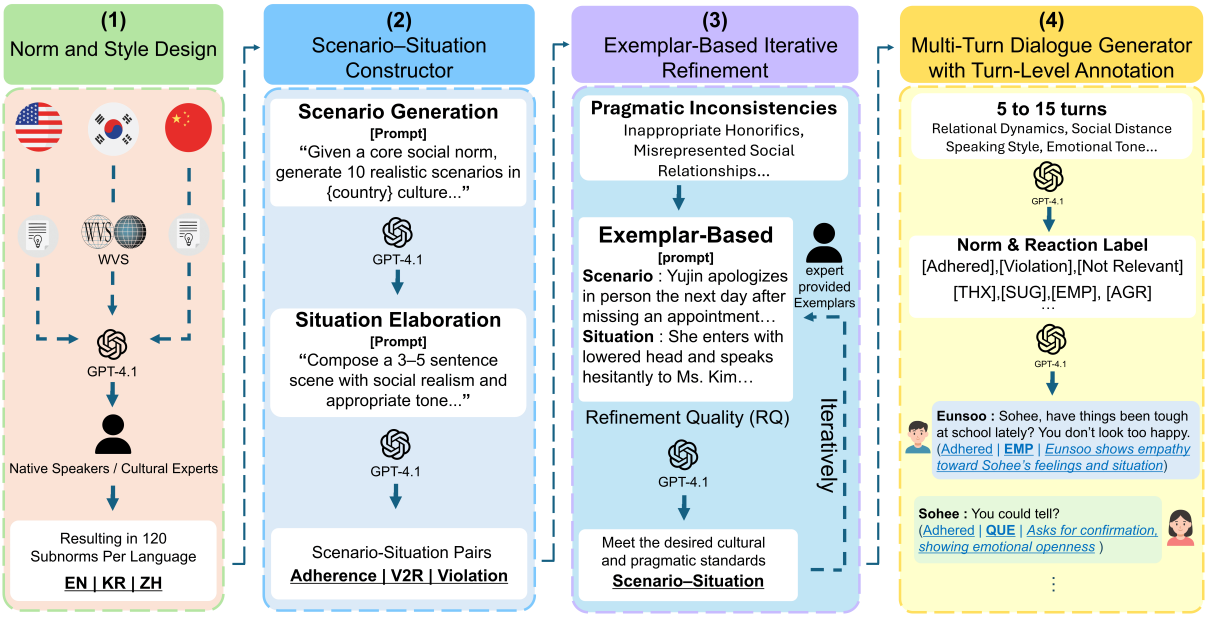

Figure 2: **NormGenesis Overview**. Our framework consists of four stages: (1) culturally grounded norm and style design, (2) scenario–situation construction across norm adherence, violation, and resolution types, (3) exemplar-based iterative refinement using semantically aligned exemplars, and (4) multi-turn dialogue generation with turn-level annotations. Each stage is evaluated and refined iteratively to ensure pragmatic consistency and cultural alignment, as described in Section 3. The RQ evaluation protocol is described in Section 4.2.

2.2 Prompt-Based and Exemplar-Guided Generation

The emergence of LLMs has enabled prompt-based synthetic data generation through techniques such as in-context learning and prompt tuning. While effective in high-resource settings, these methods often fail to capture cultural and emotional nuance in low-resource languages (Cahyawijaya et al., 2024; Pranida et al., 2025). Recent approaches (Nguyen et al., 2024; Ghosal et al., 2025) explore exemplar-based generation, but typically rely on fixed examples and prioritizes fluency over pragmatic fit. SADAS (Hua et al., 2024) also investigated norm interventions through exemplar-based remediation. However, it applies these exemplars reactively after violations occur in negotiation dialogues, remaining limited to post-hoc repair.

To address these limitations, we propose an exemplar-based iterative refinement strategy. Unlike few-shot prompting, which uses static exemplars at inference time, our method dynamically selects semantically and structurally relevant examples to guide generation. This process improves linguistic coherence and cultural alignment while enabling data bootstrapping without large-scale human annotation, particularly in low-resource languages such as Korean.

2.3 Culturally Adaptive Dialogue

As conversational AI systems expand to diverse cultural contexts, cultural adaptability becomes essential. Early multilingual datasets, such as BITOD(Lin et al., 2021) and MULTI3WOZ(Hu et al., 2023), introduced bilingual dialogues but lacked cultural annotations and failed to capture pragmatic nuance due to reliance on translation. Recent works, including CARE(Guo et al., 2025) and CULTUREPARK(Li et al., 2024), incorporate cultural preferences and simulates cross-cultural interactions with LLMs. However, these resources focus on high-resource languages and lack the fine-grained, turn-level annotations needed for culturally appropriate generation.

This gap is pronounced in underrepresented languages like Korean, where honorifics and relational pragmatics are central. To address this, we propose a multicultural framework with language-specific subnorms and turn-level annotations. Using exemplar-based iterative refinement, our method enhances cultural and pragmatic consistency, particularly in low-resource settings.

3 Method

We introduce NormGenesis, a multicultural framework for generating and refining socially grounded

Apology	Compliment
Condolence	Criticism
Empathy	Greeting
Leave-taking	Persuasion
Request	Respect
Responding to Compliments	Thanks

Table 1: Social norm categories used in this study.

dialogues in American English, Chinese, and Korean. While we broadly refer to "English" in our framework, it is important to clarify that the dataset primarily reflects U.S.-based social norms, derived from American corpora and sociocultural frameworks. This specification ensures cultural precision and avoids conflating diverse normative practices across other English-speaking contexts (e.g., British, Australian, Canadian). The framework consists of four core stages, each responsible for modeling social communication: (1) norm and style design, (2) scenario-situation constructor, (3) exemplar-based iterative refinement engine, and (4) multi-turn dialogue generator with turn-level annotation. Figure 2 illustrates the overall pipeline and stage flow. Also, the complete algorithmic workflow of our framework is provided in Appendix C.1.

3.1 Norm and Style Design

We construct a taxonomy of 12 conversational social norm categories by extending the 10 categories proposed in (Li et al., 2023) with two additional types: *Empathy* and *Respect*, motivated by prior work on dialogic functions (Stolcke et al., 2000; Bunt et al., 2020). The complete set is listed in Table 1. For each category, we define 10 culturally grounded subnorms per target language. To generate Korean-specific subnorms, we prompt an LLM with value-centric responses from the World Values Survey (WVS Wave 7, South Korea) (Haerpfer et al., 2022). English and Chinese subnorms are adapted from (Li et al., 2023) through LLM-guided alignment with the Korean outputs. All subnorms are validated by native speakers or cultural experts to ensure fluency and cultural plausibility, yielding 120 subnorms per language (Figure 2 (1)). We also define pragmatic and stylistic parameters that guide both scenario construction and dialogue generation. These include tone(formal vs. casual), honorific usage, relational distance (peer vs. hierarchical), and emotional alignment. Representative uses are shown in Table 19, with prompt templates and specifications provided in Appendix C.

3.2 Scenario-Situation Constructor

For each subnorm, we construct a scenario–situation pair consisting of: (a) a scenario that provides a concise, real-world context, and (b) a situation that expands upon the scenario with 3–5 sentences specifying relational roles, emotional states, and stylistic features such as tone and honorifics. Each instance is labeled as one of three interaction types: Norm Adherence, Norm Violation, or Violation-to-Resolution (V2R). While the first two reflect conventional norm conformity or transgression, V2R models post-violation repair strategies, capturing core aspects of interactional competence (Goffman, 2017; Feine et al., 2019). Despite its importance, V2R remains largely absent from prior norm-based dialogue datasets. To our knowledge, this is the first work to formally define and incorporate V2R into social dialogue modeling. Examples of each type are shown in Appendix D. As shown in Figure 2 (2), scenarios are first generated by prompting an LLM with a subnorm and interaction type. These are then expanded into situations via a second prompt enriched with interpersonal and emotional cues. Notably, while English outputs are generally fluent, Korean and Chinese generations often contain pragmatic inconsistencies (e.g., tone mismatch, incorrect honorifics). These issues motivate the exemplar-based refinement described in Section 3.3.

3.3 Exemplar-Based Iterative Refinement

Unlike prior approaches that rely on post-hoc filtering or manual correction after dialogue generation (Lambert et al., 2024; Occhipinti et al., 2024), our framework introduces an upstream refinement mechanism at the scenario–situation level, enabling early enforcement of cultural and pragmatic constraints. For each norm category, we manually curate a small set of high-quality exemplars that reflect culturally grounded and stylistically appropriate behaviors. Rather than using static prompts, the model retrieves semantically and structurally similar exemplars based on communicative intent, emotional tone, and discourse patterns (e.g., speaker roles and adjacency), guiding the revision process without large-scale human annotation.

To determine whether further refinement is necessary, we implement an iterative loop using our

Refinement Quality (RQ) protocol (Section 4.2), which evaluates the quality of the revised output in comparison to the original input. The model receives a triplet:

$$\left(\text{Input}_{\text{orig}},\ \text{Output}_{\text{refined}},\ \text{Score}_{\text{qual}}\right) \quad (1)$$

enabling it to assess the social and stylistic adequacy of the revision and decide whether to continue refinement. Here, $\text{Input}_{\text{orig}}$ denotes the original scenario–situation pair, $\text{Output}_{\text{refined}}$ is the revised output from exemplar-based prompting, and $\text{Score}_{\text{qual}}$ is a scalar computed via the RQ protocol.

This early integration mitigates quality bottlenecks common in generation-first pipelines and ensures consistent sociocultural alignment prior to dialogue construction. The effectiveness of our approach is illustrated in Appendix D through representative before-and-after examples across languages and norm categories, with full evaluation results reported in Section 5.1.

3.4 Multi-Turn Dialogue Generator with Turn-Level Annotation

After refinement, each scenario–situation pair is expanded into a multi-turn dialogue (5–15 turns), resulting in socially and contextually appropriate interactions. Each utterance is annotated with (a) norm adherence, (b) speaker reaction including intent and emotional state, and (c) justification for the assigned label. This structure enables fine-grained modeling of social dynamics by identifying norm compliance and explaining speaker behavior. Reaction labels, grounded in dialogue act theory (Stolcke et al., 2000), are assigned via LLM-based prompting and expert verification, as detailed in Appendix C.7. The resulting dataset supports norm reasoning and socially sensitive dialogue modeling across cultures. Representative examples of the generated dialogues and annotations are provided in Appendix D.

4 Evaluation Framework

4.1 Datasets and Experimental Conditions

Dataset Composition and Baselines. We construct a multicultural dataset in English, Chinese, and Korean, with 1,200 instances per language across three interaction types: Norm Adherence, Norm Violation, and *Violation-to-Resolution* (V2R), resulting in 10,800 instances overall. We conducted our experiments using four NVIDIA A100 GPUs over a duration of eight hours.

For the baselines, we compare against the following existing resources:

- NORMDIAL (Li et al., 2023): A bilingual English–Chinese corpus with turn-level norm adherence and violation annotations.
- SODA (Kim et al., 2023): An English corpus of socially appropriate dialogues grounded in social commonsense.

Model Configuration and Evaluation Protocol. All scenario–situation refinements and dialogue generations are conducted using GPT-4.1 (OpenAI, 2025), guided by a small set of expert-curated exemplars derived from manually revised model outputs. These exemplars reflect culturally and pragmatically appropriate responses and are used to steer iterative refinement, particularly in low-resource languages. Illustrative examples and refinement prompts are provided in Appendix C.5. Automatic evaluations are performed using GPT-4o in a zero-shot setting with metric-specific prompts as described in Appendix E.

For downstream experiments (Section 5), we fine-tune both closed and open-source language models: GPT-4o-mini as a closed model, and LLaMA-3-8B (Dubey et al., 2024), Qwen-2.5-14B, and Qwen-2.5-32B (Team, 2024) as open-source baselines. All models are trained under identical configurations for each language to ensure comparability. Human evaluations were conducted by native Chinese and Korean speakers who were selected for their linguistic fluency and cultural familiarity. To assess model performance in low-resource cultural settings, we recruited six independent graduate students (four Korean, two Chinese) as annotators. For each evaluation, we randomly sampled 100 dialogues per type, and annotators rated the outputs on fluency, relevance, and social norm adherence using a Likert-scale judgment. These human assessments were further complemented with automatic evaluations based on our proposed metrics. Detailed procedures are provided in Appendix E.

4.2 Evaluation Objectives and Design

We evaluate our framework along three axes:

1. **Refinement Quality (RQ)**: Does our refinement method improve generation quality in low-resource languages?

Criterion	Description
Norm Alignment	Adherence to the intended social norm
Language Quality	Grammaticality and fluency
Semantic Fidelity	Preservation of original intent

Table 2: Evaluation criteria used for refinement quality assessment. Scoring uses a 5-point Likert scale with both LLM and expert annotators. Detailed prompt templates for each evaluation criterion are included in Appendix E.2

Criterion	Description
Consistency	Logical flow across turns
Naturalness	Fluency and human-likeness
Relevance	Context-appropriate responses
Emotion Appropriateness	Tone aligned with context
Social Norm Appropriateness	Cultural and normative compliance
Scenario Coherence	Semantic alignment with the scenario

Table 3: Evaluation criteria used for dialogue quality (DQ). Detailed are included in Appendix E.3

2. **Dialogue Quality(DQ)**: Do generated dialogues align with norms and pragmatic expectations?

3. **Generalization Quality (GQ)**: Do our models outperform baselines in quality and human preference?

Evaluation templates for both LLM and human assessments are detailed in Appendix E, along with detailed evaluation description and guidelines.

Refinement Quality (RQ) We compare scenario-situation pairs before and after refinement in Korean and Chinese. Table 2 summarizes the three key dimensions used to evaluate refinement quality.

Dialogue Quality (DQ) We assess multi-turn dialogues along six criteria adapted from (Kim et al., 2023; Li et al., 2023). Our evaluation criteria are detailed in Table 3. LLMs and human annotators independently conduct evaluations.

Generalization Quality (GQ) We fine-tune four language models on norm-adherent dialogues from three datasets. For training, we use 1,200 English and Chinese instances from our dataset. In addition, we use 1,265 English and 1,116 Chinese instances from NORMDIAL, as well as 1,200 randomly sampled English dialogues from SODA. Models are evaluated on DAILYDIALOG (English) (Li et al., 2017) and LCCC (Chinese) (Wang et al., 2022), where each is prompted to generate five-turn continuations given benchmark dialogue contexts. We conduct A/B preference testing and human evaluation, with judgments based on social appropriateness, fluency, and overall response quality.

5 Results

We present results along the three evaluation axes defined in Section 4. Also, we further analyze the impact of our *Violation-to-Resolution* (V2R) modeling. Our framework consistently improves linguistic fluency, pragmatic coherence, and social appropriateness across all settings.

5.1 Refinement Quality (RQ)

To assess the effectiveness of exemplar-based refinement, we compare scenario-situation pairs before and after refinement in two low-resource languages: Korean and Chinese. As shown in Table 4, refinement consistently improves all evaluation dimensions across both languages. In Korean, linguistic quality improves substantially, accompanied by gains in norm alignment and semantic fidelity. Similar patterns are observed in Chinese, with a +1.32 increase in linguistic quality and near-ceiling norm alignment scores.

The refinement process is repeated until the model output is no longer selected for revision by the LLM. On average, each instance undergoes 1.2 rounds of refinement. These findings underscore the utility of our approach in enhancing fluency, coherence, and sociocultural adequacy for low-resource language generation.

To test generalizability beyond East Asian typologies, we conducted pilot refinement experiments in Malay and Urdu, two pragmatically distinct languages. Using the same evaluation setup as in Table 4, the observed improvements were consistent with those from our main experiments, supporting the robustness of our approach. Detailed results of these pilot studies are provided in Appendix F.2, which further illustrate the adaptability of our refinement strategy across diverse linguistic and cultural contexts.

5.2 Dialogue Quality (DQ)

We evaluated dialogue quality across six dimensions using both LLM- and human-based scor-

Language	Condition	Norm Align.	Linguistic Quality	Semantic Fidelity
Korean	Initial	4.577	3.589	N/A
	Refined	**4.908**	**4.910**	**4.766**
Chinese	Initial	4.855	3.603	N/A
	Refined	**4.995**	**4.926**	**4.865**

Table 4: Refinement evaluation results (RQ) in Korean and Chinese. Scores are based on a 5-point Likert scale averaged over LLM and human raters. Detailed are included in Appendix E.2

ing. Criteria are summarized in Table 3. For conciseness, we abbreviate the last three dimensions—emotional appropriateness, social norm appropriateness, and scenario–dialogue coherence—as *Emo. Approp.*, *Norm Approp.*, and *Scenario Coh.*, respectively, throughout this section.

Table 5 presents the average scores obtained from LLM-based evaluation for Adherence, V2R, and Violation scenarios in Korean, Chineses, and English. Dialogues from the *Adherence* and *V2R* categories consistently achieved high ratings (avg. > 4.9), especially for consistency, emotional appropriateness, and scenario coherence. *V2R* dialogues slightly outperformed others in emotional appropriateness and scenario–dialogue alignment, highlighting the framework's strength in modeling socially complex, repair-driven interactions. In contrast, Violation dialogues received lower scores in naturalness and emotional appropriateness across all languages, reflecting their design to capture socially inappropriate interactions.

Language	Criterion	Adherence	V2R	Violation
Korean	Consistency	**4.978**	**4.978**	2.594
	Naturalness	**5.000**	4.998	4.757
	Relevance	4.996	**5.000**	4.996
	Emo. Approp.	4.999	**5.000**	3.375
	Norm Approp.	**4.707**	3.816	1.613
	Scenario Coh.	4.988	**5.000**	4.965
Chinese	Consistency	**5.000**	4.952	1.662
	Naturalness	**4.432**	4.361	2.700
	Relevance	4.987	**5.000**	4.950
	Emo. Approp.	**5.000**	4.987	1.918
	Norm Approp.	**4.980**	3.528	1.216
	Scenario Coh.	4.896	**5.000**	4.811
English	Consistency	**5.000**	4.947	1.665
	Naturalness	**4.900**	4.623	3.381
	Relevance	**5.000**	**5.000**	4.842
	Emo. Approp.	**5.000**	4.992	2.589
	Norm Approp.	**4.982**	3.241	1.186
	Scenario Coh.	**4.994**	4.932	4.801

Table 5: Dialogue quality scores across six dimensions, evaluated on three scenario types (Adherence, *Violation-to-Resolution* (V2R), Violation) in three languages.

Human Evaluation. To validate the robustness of our LLM-based assessments and to examine generation quality in low-resource settings, we conducted a parallel human evaluation in Korean and Chinese. As shown in Table 6, human ratings exhibit patterns highly consistent with LLM scores reported in Table 5. We further quantified the alignment between LLM and human evaluations using Pearson correlation. Results indicate strong agreement across both languages, with coefficients of $r = 0.928$ (Korean) and $r = 0.945$ (Chinese). These findings confirm the reliability of our automatic evaluation protocol and support the validity of the conclusions drawn from it.

5.3 Generalization Quality (GQ)

Automatic evaluation results are summarized in Table 7. Across all models and languages, our dataset yields consistently higher preference than NORMDIAL and SODA. For instance, GPT-4o-mini was preferred in 65% (English) and 75% (Chinese) of cases over NORMDIAL, and in 65% (English) over SODA. Larger models such as Qwen-2.5-32B showed similar trends, with the strongest preference observed in Chinese. To validate these findings, we conducted blind human evaluations in Korean and Chinese (Table 8). Native speakers favored our dataset in 68% (Korean) and 77% (Chinese) of cases, closely matching LLM preferences. These results suggest that models trained on our dataset generalize better across languages and domains, generating more socially appropriate and contextually aligned responses.

We note that the untuned model occasionally produced concise, direct responses that some evaluators preferred in contexts requiring rapid apologies without nuanced emotional transitions. However, such cases were limited, and overall, our dataset achieved more than double the preference rate of the baseline. This result suggests that exemplar-guided refinement not only enhances overall quality but also provides a flexible framework for fine-grained control of response style and emotional expression, which we plan to investigate in future work.

Language	Criterion	Adherence	V2R	Violation
Korean	Consistency	4.500	**5.000**	2.500
	Naturalness	4.250	**4.625**	4.125
	Relevance	**4.750**	**4.750**	4.250
	Emo. Approp.	**4.625**	4.500	3.250
	Norm Approp.	**4.500**	4.375	1.375
	Scenario Coh.	**4.750**	4.625	4.500
Chinese	Consistency	**4.500**	4.375	1.500
	Naturalness	**4.625**	4.500	3.125
	Relevance	**5.000**	4.750	4.500
	Emo. Approp.	4.500	**5.000**	1.375
	Norm Approp.	**4.750**	4.500	1.500
	Scenario Coh.	**4.750**	4.625	4.125

Table 6: Human evaluation results across dialogue types (Adherence, *Violation-to-Resolution* (V2R), Violation) and six dimensions. Scores are based on a 5-point Likert scale.

Model	Language	Ours vs. NormDial	Ours vs. SODA
GPT-4o-mini	English	**65%**	**65%**
	Chinese	**75%**	N/A
LLaMA-3-8B	English	**65%**	**59%**
Qwen-2.5-14B	English	**51%**	**61%**
	Chinese	**71%**	N/A
Qwen-2.5-32B	English	**56%**	**62%**
	Chinese	**79%**	N/A

Table 7: A/B test results comparing preference for models trained on our dataset versus NormDial and SODA. Each value represents the percentage of times responses from models trained on *our dataset* were preferred by annotators.

5.4 Effect of Violation-to-Resolution

To assess the impact of *Violation-to-Resolution* (V2R) training on norm-sensitive generation, we conduct a focused comparison using PROSOCIALDIALOG (Kim et al., 2022), a benchmark for ethically challenging scenarios. We fine-tune two GPT-4o-mini models under comparable conditions: one on the full NormDial dataset and one on an equal-sized subset of our data, including three types of our datasets. Both models are prompted with 100 norm-violating contexts, each requiring a five-turn continuation.

Blind A/B human evaluations show that the V2R-augmented model is preferred in 82% of cases (Table 9), with annotators consistently favoring its empathy, contextual fit, and ability to model norm repair. These findings underscore the utility of V2R as a training signal for enhancing pragmatic competence in ethically sensitive dialogue and support its integration into norm-grounded generation frameworks.

Language	Ours	Untuned GPT-4o-mini	NormDial
Korean	**68%**	32%	N/A
Chinese	**77%**	5%	18%

Table 8: Human preference results comparing our dataset with untuned GPT-4o-mini and NormDial. Evaluations were conducted under blind conditions with native speakers. Korean evaluation compares against untuned GPT-4o-mini, while Chinese includes baseline.

6 Discussion

Limitations of Prompt-Based Generation in Low-Resource Contexts. Our refinement framework is motivated in part by the limitations of prior prompt-based approaches to social norm generation (Li et al., 2023; Zhan et al., 2024). These methods typically rely on static prompts with minimal norm signals, placing the burden of generation entirely on the language model. While this approach may yield fluent and contextually appropriate responses in high-resource languages such as English, it often results in pragmatic failures in low-resource settings like Korean and Chinese.

In Korean, generated dialogues frequently exhibit lexical redundancy (e.g., repeated expressions) and tone mismatches (e.g., informal apologies in formal contexts). In Chinese, issues include unnatural phrasing, exaggerated emotional responses, repetitive honorifics, and inconsistent tone from register mixing. These limitations underscore the difficulty of capturing fine-grained sociocultural norms through prompt-only methods.

To mitigate these issues, we introduce a refinement framework to improve fluency and norm alignment in low-resource settings. Additional examples appear in Appendix F.1.

Early Refinement for Sociocultural Alignment. As shown in Section 5.1, even a small set of high-quality exemplars at this stage improves fluency and norm alignment in low-resource languages. Prior to refinement, we conducted a comparative analysis between model-generated outputs and native-authored revisions, which revealed recurring issues such as overuse of formulaic expressions, limited gesture variety, register–context mismatch, and weakened hierarchical cues. A key insight from this analysis is the distinction between surface accuracy and cultural appropriateness: model outputs may be grammatically and semantically correct, yet fail to include ritualistic or affective elements (e.g., apologies, condolences) that are

Training Data	Preference (%)
Ours (with V2R)	**82**
NormDial	18

Table 9: Human preference results on PROSOCIALDIALOG. Models trained on our dataset with V2R significantly outperformed those trained on NormDial.

Category	Strategy / Sequence	English	Chinese	Korean
Strategy (%)	Apology (A)	98.7	93.0	92.6
	Explanation (X)	91.1	88.1	86.3
	Empathy (E)	90.6	62.0	89.2
	Compensation (C)	82.8	79.3	72.4
	Humor (H)	12.6	7.9	12.0
Top Sequence	Sequence	X → A → C	A → X → C	E → A → X
	Frequency (%)	33	29	32

Table 10: Strategy usage rates and most frequent recovery sequences in V2R dialogues across American English, Chinese, and Korean.

essential to pragmatic expectations. These shortcomings were particularly salient in low-resource settings and are illustrated in Appendix F.1.

When optimized early in the generation pipeline, scenario–situation pairs provide strong social cues that guide the construction of coherent and culturally aligned dialogues. This early-stage refinement approach also aligns with recent works in controllable generation and structured planning (Moryossef et al., 2019; Rashkin et al., 2021), which emphasize the role of explicit context modeling in coherence and goal alignment. Since social norms are inherently entangled with relational roles, power dynamics, and situational contexts, refining situational priors ensures that pragmatic and culturally appropriate behaviors emerge naturally. Our refinement stage thus functions not as a post hoc correction layer, but as a core mechanism for embedding sociocultural alignment into the generative process.

Qualitative Insights into Cross-Linguistic Repair Strategies Our qualitative analysis of Violation-to-Resolution (V2R) dialogues shows that the proposed framework effectively models both universal and culture-specific repair strategies. As summarized in Table 10, Apology and Explanation dominate across English, Chinese, and Korean. Still, sequencing patterns diverge in culturally meaningful ways: English dialogues most often follow Explanation → Apology → Compensation, Chinese dialogues Apology → Explanation → Compensation, and Korean dialogues Empathy → Apology → Explanation. These results, informed by established taxonomies in apology and politeness research (Radu et al., 2019; Zhang and Wang, 2024), validate the framework's ability to capture nuanced cross-linguistic variation. Notably, over 85% of V2R dialogues employed multi-step recovery, confirming that socially coherent repair rarely occurs through a single act. By integrating exemplar-guided refinement, our approach models these layered dynamics, bridging computational dialogue generation with sociolinguistic insights. These findings highlight the dual value of the V2R paradigm: providing a scalable schema for realistic dialogue repair and serving as a diagnostic lens for cultural variation. This adaptability also points to promising directions for extending NormGenesis to typologically diverse languages (e.g., Arabic, Swahili, Hindi), where divergent pragmatic systems pose additional challenges.

7 Conclusion

We present NormGenesis, a multicultural framework for generating and refining socially grounded dialogues in English, Chinese, and Korean. To address cultural and pragmatic limitations of existing dialogue systems, particularly in low-resource settings, we introduce an exemplar-based iterative refinement applied at the scenario-situation level. This upstream refinement design enables early alignment with linguistic, emotional, and sociocultural expectations, reducing generation errors before full dialogue synthesis. We further propose a novel dialogue type, *Violation-to-Resolution* (V2R), which models the recovery process following norm violations through repair strategies. V2R facilitates more realistic and context-sensitive modeling of social interaction dynamics. Our experimental results show that V2R not only improves pragmatic competence in ethically sensitive scenarios but also enhances generalization across languages and domains. Through comprehensive human and LLM-based evaluations, we demonstrate that NormGenesis consistently outperforms existing datasets such as NORMDIAL and SODA across multiple dimensions, including norm alignment, emotional coherence, and repair quality. By integrating linguistically and culturally diverse norms, fine-grained turn-level annotations, and structured refinement, NormGenesis provides a scalable and robust foundation for norm-aware dialogue modeling in multilingual and multicultural contexts.

Limitations

While NormGenesis achieves strong performance across linguistic, emotional, and social dimensions, we acknowledge several limitations that open key directions for future work.

Language Coverage. Our framework currently supports only English, Chinese, and Korean. While these languages span a spectrum of resource availability and cultural characteristics, the framework does not address the full diversity of global languages and interactional norms. Expanding NormGenesis to additional languages—particularly those with limited computational resources or distinct social conventions (e.g., Arabic, Hindi, Swahili)—remains a key avenue for future work.

To test generalizability beyond East Asian typologies, we conducted pilot refinement in Malay and Urdu, two pragmatically distinct languages. The results closely aligned with those from our main experiments, demonstrating the framework's capacity to generalize to typologically and culturally diverse settings. Detailed results are presented in Appendix F.2. Building on this, we plan to extend NormGenesis to further low-resource languages, including Arabic and Swahili.

Exemplar Scalability. The iterative refinement process relies on a small number of manually revised exemplars. Although this approach is more scalable than full human annotation, scaling to a large number of new norms or domains could still be resource-intensive. To mitigate this, future work will explore active learning for efficient exemplar selection, a structured norm-centric repository for retrieval-based reuse, and clustering of culturally aligned regions (e.g., via World Values Survey) to enable exemplar transfer. These strategies aim to improve coverage and ensure scalable, high-quality refinement.

Evaluation and Subjectivity. Evaluating social norm adherence and conversational appropriateness inevitably involves subjectivity and cultural bias, especially across diverse sociolinguistic contexts. To mitigate this, we adopted three safeguards: (1) detailed rubrics assessing norm alignment, fluency, and emotional appropriateness (Appendix E); (2) native speaker annotators with cultural expertise; and (3) cross-review by multiple experts to offset exemplar-induced bias. While these measures and high inter-annotator agreement enhance reliability, further reducing cultural subjectivity remains an open challenge. Future research could explore culturally calibrated evaluation protocols or leverage LLM-based evaluators fine-tuned on localized criteria.

Norm Evolution. Social norms are dynamic and shift across time, communities, and platforms. Our taxonomy provides a structured but time-bounded snapshot. Systematically tracking and modeling the evolution of norms over time and across social contexts will be important for adaptive and future-proof dialogue systems. Future work should systematically track norm shifts using longitudinal corpora or real-time social data, enabling adaptive norm modeling for evolving conversational environments.

Ethical Considerations

NormGenesis aims to advance the development of culturally adaptive and socially competent dialogue agents by modeling nuanced social norms across English, Chinese, and Korean. However, several ethical considerations warrant discussion.

Intended Use and Misuse. Our dataset is designed for training dialogue systems to generate socially appropriate, norm-aware responses. As with any dataset containing norm violations and repair strategies, there is a risk that malicious users could exploit the resource to train agents that generate inappropriate or harmful utterances. We urge the community to use the dataset solely for prosocial, culturally sensitive, and norm-aligned conversational AI research.

Cultural Scope and Generalizability. NormGenesis reflects culturally salient behaviors as of the time of data collection. Despite extensive native speaker review and expert refinement, social norms are inherently dynamic and context-dependent. Caution is advised when applying the resource to new languages, regions, or changing societal contexts, as some outputs may not generalize beyond the represented cultures.

Annotation Subjectivity and Bias. All sub-norms and dialogues are reviewed or annotated by cultural experts and native speakers, but the interpretation of social appropriateness and emotional tone involves subjective judgment. While inter-annotator agreement is high, some bias may persist, especially for edge cases or rapidly chang-

ing norms. Broadening annotator diversity and incorporating community feedback may mitigate these effects.

Dataset Balance and Representation. NormGenesis includes a diverse range of norm-adhering, violating, and Violation-to-Resolution (V2R) dialogues. However, its scenario coverage is not exhaustive and may reflect existing cultural, demographic, or linguistic biases. Supplementing NormGenesis with additional resources is encouraged to ensure robust and contextually sensitive conversational agents.

Potential for Negative Outcomes. While the V2R paradigm models constructive responses to norm violations, dialogue agents trained on these data should not be used for critical decision-making or sensitive applications (e.g., counseling, legal advice) without careful human oversight. The framework is intended to support research and development in social dialogue modeling, not to replace professional judgment. All code, data, and annotation guidelines will be released publicly upon acceptance, promoting transparency, reproducibility, and responsible community use.

Acknowledgments

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2025-RS-2020-II201789), and the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2025-RS-2023-00254592) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation, second edition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 549–558, Marseille, France. European Language Resources Association.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Jon Elster. 2006. Fairness and norms. *Social Research: An International Quarterly*, 73(2):365–376.

Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A taxonomy of social cues for conversational agents. *International Journal of human-computer studies*, 132:138–161.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670.

Soumya Suvra Ghosal, Soumyabrata Pal, Koyel Mukherjee, and Dinesh Manocha. 2025. PromptRefine: Enhancing few-shot performance on low-resource Indic languages with example selection from related example banks. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 351–365, Albuquerque, New Mexico. Association for Computational Linguistics.

Erving Goffman. 2017. *Interaction ritual: Essays in face-to-face behavior*. Routledge.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. Care: Aligning language models for regional cultural awareness. *Preprint*, arXiv:2504.05154.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jose Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen, editors. 2022. *World Values Survey: Round Seven – Country-Pooled Datafile Version 6.0*. JD Systems Institute & WVSA Secretariat, Madrid, Spain & Vienna, Austria.

Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023. Multi3woz: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems. *Preprint*, arXiv:2307.14031.

Yuncheng Hua, Lizhen Qu, and Reza Haf. 2024. Assistive large language model agents for socially-aware negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8047–8074, Miami, Florida, USA. Association for Computational Linguistics.

Seongbo Jang, Seonghyeon Lee, and Hwanjo Yu. 2024. KoDialogBench: Evaluating conversational understanding of language models with Korean dialogue

benchmark. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9905–9925, Torino, Italia. ELRA and ICCL.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, and 1 others. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029.

Nathan Lambert, Hailey Schoelkopf, Aaron Gokaslan, Luca Soldaini, Valentina Pyatkin, and Louis Castricato. 2024. Self-directed synthetic dialogues and revisions technical report. *arXiv preprint arXiv:2407.18421*.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024. KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11177–11213, Bangkok, Thailand. Association for Computational Linguistics.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024. Culturepark: Boosting cross-cultural understanding in large language models. *Preprint*, arXiv:2405.15145.

Oliver Li, Mallika Subramanian, Arkadiy Saakyan, CH-Wang Sky, and Smaranda Muresan. 2023. Normdial: A comparable bilingual synthetic dialog dataset for modeling social norm adherence and violation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15732–15744.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

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

Bertram F Malle, Steve Guglielmo, and Andrew E Monroe. 2014. A theory of blame. *Psychological Inquiry*, 25(2):147–186.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2267–2277.

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516.

Daniela Occhipinti, Michele Marchi, Irene Mondella, Huiyuan Lai, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2024. Fine-tuning with HED-IT: The impact of human post-editing for dialogical language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11892–11907, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2025. Openai models documentation. Accessed: 2025-04-01.

Salsabila Zahirah Pranida, Rifo Ahmad Genadi, and Fajri Koto. 2025. Synthetic data generation for culturally nuanced commonsense reasoning in low-resource languages. *Preprint*, arXiv:2502.12932.

Alexandru Gabriel Radu, Denni Arli, Jiraporn Surachartkumtonkun, Scott Weaven, and Owen Wright. 2019. Empathy and apology: The effectiveness of recovery strategies. *Marketing Intelligence & Planning*, 37(4):358–371.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. 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 704–718.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2022. A large-scale chinese short-text conversation dataset. *Preprint*, arXiv:2008.03946.

Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds, Suraj Sharma, Shilin Qu, Linhao Luo, Ingrid Zukerman, Lay-Ki Soon, Zhaleh Semnani Azad, and Reza Haf. 2024. RENOVI: A benchmark towards remediating norm violations in socio-cultural conversations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3104–3117, Mexico City, Mexico. Association for Computational Linguistics.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, and 1 others. 2023. Socialdial: A benchmark for socially-aware dialogue systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2712–2722.

Junran Zhang and Nan Wang. 2024. The impact of humorous apology expression on consumer forgiveness and trust rebuilding after trust violations.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, Junhao Chen, and Tianming Liu. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research. *Preprint*, arXiv:2412.04497.

A Dataset Overview

A.1 Dataset Composition Summary

Our dataset includes multicultural social norm dialogues across 12 categories: *Apology, Compliment, Condolence, Criticism, Empathy, Greeting, Leave-taking, Persuasion, Request, Respect, Responding to Compliments, Thanks*. For each category, we define 10 subnorms per language (English, Korean, Chinese), resulting in 120 subnorms per language.

- **Total Subnorms**: 360 (120 per language)
- **Total Scenario–Situation Pairs**: 10,800 (3 types × 3 languages × 1,200 each)
- **Total Dialogues**: 10,800 (1 per instance)
- **Total Average Turn**: 11.91 Turn.

Adherence

Lang	Cat	Sub	Scen	Situ	Dial	AvgT
EN	12	120	1200	1200	1200	10.56
KR	12	120	1200	1200	1200	10.21
ZH	12	120	1200	1200	1200	10.74

Table 11: Dataset statistics for Adherence

Violation-to-Resolution (V2R)

Lang	Cat	Sub	Scen	Situ	Dial	AvgT
EN	12	120	1200	1200	1200	16.06
KR	12	120	1200	1200	1200	12.56
ZH	12	120	1200	1200	1200	13.49

Table 12: Dataset statistics for violation-to-resolution (V2R)

A.2 Norm-Type Statistics

To provide an overview of the social norm taxonomy introduced in this study, Table 11, 12, and 13 present detailed summary statistics for the Adherence, Violation-to-Resolution (V2R), and Violation categories, respectively. For brevity, column headers are abbreviated as follows: Lang = Target Language, Cat = Norm Category, Sub = Subnorm, Scen = Scenario, Situ = Situation, Dial = Dialogue, AvgT = Average Turn Count per dialogue. These tables summarize the scale and structure of our dataset across all categories and languages.

A.3 Subnorm Coverage and Cultural Examples

Each of the 12 norm categories comprises 10 culturally grounded subnorms per language, resulting in a total of 360 subnorm definitions with aligned examples across English, Korean, and Chinese. These examples serve as reference points for scenario generation and support culturally appropriate dialogue construction in each language. Full examples are provided in Appendix C.

A.4 Instance Structure

Each dialogue instance in the dataset is composed of the following stages:

- **Norm Category & Subnorm:** A high-level social norm category and its culturally grounded subnorm definition.
- **Scenario:** A brief 1–2 sentence description outlining the situational context in which the norm is relevant.

Violation						
Lang	**Cat**	**Sub**	**Scen**	**Situ**	**Dial**	**AvgT**
EN	12	120	1200	1200	1200	11.59
KR	12	120	1200	1200	1200	11.17
ZH	12	120	1200	1200	1200	10.82

Table 13: Dataset statistics for violation

- **Situation:** A 3–5 sentence elaboration of the scenario that specifies tone, interpersonal relationship, and emotional cues to guide dialogue generation.
- **Dialogue:** A multi-turn conversation consisting of 5 to 15 turns that reflects the defined norm and situational context.
- **Annotations:** Turn-level labels for social norm adherence (e.g., Adherence, Violated, V2R) and speaker reactions (e.g., Apology, Empathy, Agreement), enabling fine-grained evaluation of social behavior and pragmatic intent.

This structured format enables controlled generation and fine-grained annotation of socially grounded dialogues across multiple languages.

A.5 Language Balance and Complexity Metric

To ensure cultural and linguistic balance, each language is equally represented across all norm types. While average token length is commonly used to measure dialogue complexity, we omit it here due to tokenizer variations across models. Instead, we report average dialogue turns as a consistent and model-independent proxy for complexity.

B Annotation Schema

This appendix describes the schema used for turn-level annotation of generated dialogues. Each utterance is annotated with a communicative function label drawn from a predefined set of categories, facilitating fine-grained analysis of pragmatic intent and interactional structure across cultural contexts.

B.1 Label Set Definition

We employ a set of 11 functional dialogue act labels to annotate each turn, as summarized in Table 14. These labels capture core social and communicative functions, including acknowledgment, apology, suggestion, justification, and other norm-relevant speaker intentions.

B.2 Annotation Usage

Each generated dialogue is annotated at the turn level. Given a dialogue $D = u_1, ..., u_n$ consisting of n utterances, we assign a label $y_i \in \mathcal{Y}$ to each turn u_i using an automatic annotation function f_{annotate}. This annotation framework supports comparative analysis of pragmatic behavior across the following dimensions:

- **Norm Type:** Adherence, Violation, Violation-to-Resolution (V2R)
- **Cultural Context:** English (EN), Korean (KR), Chinese (ZH)
- **Speaker Role:** e.g., subordinate vs. authority figure

This enables a structured investigation of how social norms and communicative functions vary across languages and roles.

B.3 Annotation Example

Appendix C.7 details the prompt formulations used to perform turn-level annotation. For illustration, Figure 7 presents a turn-level annotation example for the "Adherence" category in English.

C Generation & Refinement Prompt Templates

This appendix presents the full set of prompt templates used throughout the multilingual dialogue generation and refinement pipeline. We first outline the overall algorithmic workflow, which defines the step-by-step procedures for constructing culturally grounded dialogues. We then provide task-specific prompt templates corresponding to each stage of the pipeline, designed to ensure consistency, linguistic fluency, and norm alignment across languages.

C.1 Algorithm of Our Framework

Algorithm 1 outlines the complete pipeline used to construct our multilingual, norm-grounded dialogue dataset. The framework consists of four main stages:

Step 1: Social Norm Construction. For each language l and social norm category c, we generate a set of subnorms $\mathcal{N}_{l,c}$ that encode fine-grained, culturally grounded expectations. These subnorms serve as foundational inputs for subsequent stages of scenario and dialogue generation.

Label	Description
ACK (Acknowledgment)	Explicit acknowledgment that the speaker has heard or understood the interlocutor's statement.
AGR (Agreement)	Expressing agreement or alignment with the other person's opinion or position.
DIS (Disagreement / Refusal)	Expressing disagreement or rejecting a suggestion.
APO (Apology)	Expressing regret or remorse for one's mistake or inconvenience caused.
THX (Gratitude)	Expressing appreciation for help, kindness, or praise.
EMP (Empathy / Support)	Emotionally validating or supporting the interlocutor's feelings or situation.
JUS (Justification)	Offering an explanation or excuse to justify one's actions or mistakes.
SUG (Suggestion / Advice)	Proposing a solution or sharing an opinion to help resolve an issue.
QUE (Question / Clarification Request)	Asking a question to seek further explanation or information.
CRT (Criticism)	Pointing out problems or expressing negative evaluations of the other person's actions or statements.
N/A (Not Applicable)	Used when the utterance does not clearly fall into any of the defined categories above. Typically applies to greetings, topic transitions, structural openers, or filler phrases.

Table 14: Turn-level annotation labels and descriptions

Step 2: Scenario–Situation Generation. For each subnorm n and dialogue type t (e.g., Adherence, Violation, Violation-to-Resolution), we construct a scenario $S_{l,n,t}$ that outlines the relevant social context. This is followed by a situation $T_{l,n,t}$, which elaborates on tone, relationship, and emotional dynamics to constrain downstream dialogue construction.

Step 3: Exemplar-Based Iterative Refinement. To ensure cultural and pragmatic fidelity, an expert manually refines a single exemplar $(S_{l,n,t}, T_{l,n,t})$ per subnorm. This exemplar guides LLM-based refinement of structurally or semantically similar pairs. The refinement process is repeated until the quality score $Q_{l,n,t}$ exceeds a predefined threshold, ensuring consistency across instances.

Step 4: Dialogue Generation and Annotation. Each refined scenario–situation pair $(S_{l,n,t}, T_{l,n,t})$ is used to generate a multi-turn dialogue $D_{l,n,t}$, conditioned on the norm category, subnorm, scenario, and situation. All dialogue turns are annotated with norm and reaction labels, resulting in a fully labeled dataset $\mathcal{D}$ suitable for training and evaluation.

Summary. This structured pipeline enables the generation of culturally faithful, pragmatically coherent, and richly annotated dialogues. It supports multilingual benchmarking and serves as a scalable foundation for norm-aware dialogue modeling.

C.2 Subnorm Generation Prompt

To generate culturally grounded subnorms for the 12 norm categories defined in Section 3.1, we construct 10 subnorms per category for each target language, yielding 360 subnorms in total. For Korean, where prior work on normative dialogue modeling is limited, we leverage sociocultural value indicators from the World Values Survey (WVS Wave 7, South Korea) as illustrated in Table 15. For English and Chinese, we adopt the subnorm definitions from (Li et al., 2023) as semantic anchors. To ensure cross-cultural consistency, we design prompts that adapt these definitions to align with the Korean-derived subnorms. Table 16 illustrates representative prompt templates used for this alignment procedure. As a result, we design a total of 360 subnorms across the 12 categories and three languages, with illustrative examples shown in Figure 3. This protocol ensures cross-linguistic consistency while preserving cultural specificity and serves as the foundation for downstream scenario construction and dialogue generation.

C.3 Scenario Generation Prompt

To generate culturally grounded data, we first construct concise yet diverse scenarios aligned with a given subnorm within each social norm category. For each target language (English, Korean, Chinese), we prompt the model to produce 10 distinct and contextually appropriate scenarios per subnorm. The prompts are designed to ensure that

Section	Content
Target Language	< Korean >
Parameter	< WVS Responses >
Instruction	You are a culturally aware assistant with deep knowledge of Korean social norms and communication practices. Given a specific social norm category (e.g., Apology, Empathy), your task is to generate 10 Korean-specific conversational subnorms that reflect the core values and expectations found in Korean society. These values are derived from nationally representative survey data, including interpersonal relationships, formality, group harmony, respect for authority, and emotional expression. Please follow these instructions when generating the subnorms: 1. Ensure that each subnorm reflects Korean cultural values and not generic or universal norms. 2. Specify the context in which the norm should be applied (e.g., school, workplace, with elders). 3. Include verbal evidence (i.e., example phrases in Korean) that would signal adherence to the subnorm in dialogue. 4. The subnorms should be actionable and observable in conversation.
Format	Subnorm 1 Subnorm 2 ... Subnorm 10

Table 15: Prompt for korean subnorm generation.

the resulting scenarios are socially plausible and culturally relevant. As shown in Table 17, these prompts provide structured guidance for consistent and culturally sensitive scenario generation.

C.4 Situation Elaboration Prompt

Building on each generated scenario, this prompt instructs the model to produce a realistic and emotionally coherent situation in 3–5 sentences. The generated text is expected to reflect culturally appropriate tone, interpersonal dynamics, and narrative plausibility. These situational descriptions serve as the contextual foundation for downstream dialogue generation. A representative example is provided in Table 18.

C.5 Exemplar-Based Refinement Prompt

To improve linguistic quality and pragmatic fidelity, particularly in low-resource languages, we employ a one-shot refinement strategy. The model is given an expert-curated Scenario–Situation pair as an exemplar and instructed to revise a batch of initial outputs to match the demonstrated level of cultural appropriateness, contextual richness, and fluency. Table 19 illustrates the structure and usage of the refinement prompt.

C.6 Multi-Turn Dialogue Generation Prompt

Given a refined scenario and situation, this prompt guides the generation of a natural, coherent multi-turn dialogue that adheres to the specified social norm. Each dialogue comprises 5 to 15 turns and is expected to reflect appropriate cultural tone, relational dynamics, and norm-conforming behavior in a realistic conversational format. An illustrative example is provided in Table 20.

C.7 Turn-Level Annotation Prompt

To assess norm adherence and communicative function at the utterance level, this prompt guides the annotation of each dialogue turn with: (1) a norm label (Adherence, Violation, Not Relevant), (2) a communicative function tag (e.g., APO, ACK, THX), and (3) a brief justification.

This structured annotation facilitates consistent, turn-level analysis of social behaviors and pragmatic functions across cultures. An example is shown in Table 21.

Section	Content
Target Language	`< Chinese | English >`
Parameter	`< Korean subnorm | Normdial subnorm >`
Instruction	You are a culturally adaptive assistant with knowledge of conversational norms across multiple languages. Provided with a Korean subnorm reflecting a specific cultural value (e.g., how to apologize, express empathy, or show respect), your task is to generate a corresponding conversational subnorm in (Chinese | English) that matches the Korean subnorm in meaning and function. Use the following instructions: 1. Ensure the subnorm reflects the target language's cultural norms (i.e., Chinese or English), not just a literal translation of the Korean input. 2. Specify the context in which the norm should be applied (e.g., school, workplace, with elders). 3. Include verbal evidence (i.e., example phrases in Chinese or English) that would signal adherence to the subnorm in dialogue. 4. The subnorms should be actionable and observable in conversation.
Format	Subnorm 1 Subnorm 2 ... Subnorm 10

Table 16: Prompt for chinese and english subnorm generation.

D Generation & Refinement Examples

This appendix presents full examples of generated and refined outputs for each social norm type—*Adherence*, *Violation*, and *Violation-to-Resolution (V2R)*—across English (EN), Korean (KR), and Chinese (ZH). For each case, we provide the subnorm definition, the Scenario–Situation pair before and after refinement, the corresponding dialogue, and representative turn-level annotations.

To facilitate narrative understanding and cross-cultural comparison, we include aligned figure references for each norm type and language. These visualizations illustrate the transformation process and demonstrate how refinement enhances linguistic fluency, pragmatic appropriateness, and sociocultural alignment. All examples are drawn from the *Apology* category.

D.1 Adherence Examples

This section presents example dialogues that adhere to social norms from the initial generation stage, across English, Korean, and Chinese. Although these dialogues already demonstrate norm-conforming behavior, we apply exemplar-based refinement to enhance fluency, tonal consistency, and cultural appropriateness. In high-resource languages like English, model outputs tend to exhibit strong cohesion and emotional clarity even before refinement (Figure 4. The initial dialogue demonstrates contextual awareness and sincerity with minimal pragmatic inconsistencies. In contrast, the Korean and Chinese examples reveal more pronounced cultural deviations. For instance, Korean outputs often lack the deference and softened phrasing expected in hierarchical contexts (Figure 5, while Chinese dialogues may omit expressions of empathy or communal responsibility crucial in professional settings (Figure 6. Post-refinement, all examples show marked improvements in pragmatic subtlety, such as calibrated tone, culturally appropriate honorifics, and clearer interpersonal alignment, yielding more realistic and socially congruent conversations. Turn-level annotation is subsequently applied to each dialogue, capturing both norm adherence and speaker intent at the utterance level. An annotated example illustrating this process is shown in Figure 7.

D.2 Violation-to-Resolution(V2R) Examples

This section presents dialogues that initially violate social norms but subsequently demonstrate conversational repair through culturally appropri-

Section	Content
Target Language	`< Korean | English | Chinese >`
Category	`< Apology >`
Instruction	Based on the above `Subnorm` for the given `Category`, generate 10 distinct and concise scenarios that could naturally occur in a similar social context within `Country` culture. Please ensure realism and cultural grounding. Use names and honorifics commonly used in the specified country.
Input Format	`Category, Subnorm, Instruction, Type`
Output Format	Scenario 1 Scenario 2 ... Scenario 10

Table 17: Prompt for scenario generation.

Section	Content
Target Language	`< Korean | English | Chinese >`
Category	`< Apology >`
Instruction	Generate a culturally plausible Situation in 3–5 sentences. 1. Use realistic names and honorifics. 2. Ensure emotional coherence. 3. Depict through action/dialogue, not explanation.
Input Format	`Category, Subnorm, Scenario, Instruction`
Output Format	Situation 1 Situation 2 ... Situation 10

Table 18: Prompt for situation elaboration.

ate resolution strategies. These examples reflect how speakers can realign interactions with social expectations by acknowledging fault, expressing remorse, and adopting conciliatory tones—core mechanisms of the Violation-to-Resolution (V2R) paradigm. Across English, Korean, and Chinese, pre-refinement dialogues exhibit typical pragmatic violations: abrupt interruptions, deflection of responsibility, or insufficient emotional engagement. For instance, English outputs show initial breaches in conversational protocol (e.g., interrupting a professor, as illustrated in Figure 8), while Korean and Chinese versions reveal culturally incongruent justification strategies or failures to express appropriate deference (Figure 9, 10). Through exemplar-guided refinement, these dialogues are revised to incorporate explicit acknowledgments of fault, context-sensitive apologies, and relational mitigation techniques. The resulting interactions better conform to cultural norms governing hierarchy, face management, and affective alignment. Turn-level annotation is applied to each refined dialogue to encode both norm adherence and speaker intent systematically. A representative annotation example illustrating this process is provided in Figure 11.

D.3 Violation Examples

This section presents dialogues intentionally constructed to illustrate clear violations of social norms without subsequent repair. These examples are designed to expose pragmatic failures, such as disregard for authority, lack of emotional engagement, or avoidance of responsibility, and to demonstrate their effects on interpersonal dynamics.

The selected instances span English, Korean, and

Section	Content
Instruction	Given a naive Scenario–Situation pair and an expert-refined version, rewrite all other naive inputs to match the expert style. Do not shorten content. Ensure cultural richness and tone. Cover a wide range of settings and relationships. Avoid repetition.
Input Format	`Category, Subnorm, 9 Naive Scenarios, 9 Naive Situations, Exemplar`
Output Format	Rewritten Scenario 1 Rewritten Situation 1 ... Rewritten Scenario 9 Rewritten Situation 9

Table 19: Prompt for exemplar-based scenario–situation refinement.

Section	Content
Input stages	`Category, Subnorm, Scenario, Situation`
Instruction	Generate a natural and realistic dialogue between two speakers reflecting the Scenario and Situation above. Write in English. Format each turn with speaker names. Dialogue should be 5–15 turns long.

Output Format	`Name: line of dialogue`
	`Name: line of dialogue`
	`... [END]`

Table 20: Prompt for multi-turn dialogue generation.

Chinese, each reflecting culture-specific patterns of norm deviation. In English, violations appear as dismissive behavior toward institutional figures (for example, laughing off a misstep with a principal, as illustrated in Figure 12). In Korean, pragmatic breakdowns occur in professional settings where the speaker avoids apologizing by offering repeated justifications, diverging from culturally expected norms of deference (Figure 13). Chinese examples show minimal accountability and disengaged behavior, violating expectations of relational harmony and respect in hierarchical contexts (Figure 14).

These dialogues are preserved in their original form to maintain the narrative dissonance they introduce. To support structured analysis, each turn is annotated with norm adherence and communicative function labels. A representative annotation example is shown in Figure 15.

E Evaluation Setup

Human Evaluation and Annotator Agreement As described in Section 4 and Section 5, we conduct four distinct evaluation tasks to assess the quality, coherence, and norm alignment of generated outputs. Each task is guided by a dedicated prompt tailored to its respective objective. The complete prompt formulations used for generation and evaluation are provided in the appendix.

To assess model performance in low-resource cultural settings, we recruited six graduate students as human annotators, all of whom were independent from the research team. Among them, four were native Korean speakers and two were native Chinese speakers, each with over ten years of immersion in their respective cultural environments. Annotators rated model outputs using Likert-scale judgments across multiple evaluation dimensions, including fluency, relevance, and social norm adherence. All annotators were compensated fairly in accordance with ethical research guidelines.

Inter-annotator agreement was calculated using Krippendorff's Alpha (α) within each language group. Korean annotators ($n = 4$) demonstrated strong agreement ($\alpha = 0.81$), and Chinese annotators ($n = 2$) achieved substantial reliability ($\alpha = 0.72$), indicating internal consistency and

Section	Content
Instruction	For each dialogue turn, label: (1) Norm-level adherence (Adherence / Violation / Not Relevant), (2) Speaker's reaction label (e.g., APO, ACK, AGR), and (3) Explanation for label.
Label Format	`Role | Norm Label | Reaction Label | Explanation`
Reaction Labels	APO: Apology, ACK: Acknowledgment, AGR: Agreement, DIS: Disagreement, THX: Thanks, EMP: Empathy, JUS: Justification, SUG: Suggestion, QUE: Question, CRT: Criticism, N/A: Not applicable

Table 21: Prompt for turn-level annotation.

cultural coherence in the evaluation process.

IRB Information This study involved human participants for data refinement and evaluation tasks conducted in South Korea. All data used were synthetic, generated by large language models, and did not include any personally identifiable information. Human experts manually refined the corpus, and recruited annotators performed human evaluation under informed instructions. No crowdworker IDs or personal data were collected or stored, and all annotations were submitted anonymously. Participants were provided with fair compensation according to the guidelines specified in the task description. Based on national research ethics standards and in alignment with U.S. federal regulation 45 CFR 46, this study qualifies as exempt from formal IRB review.

E.1 Dialogue Dataset & Language Models Descriptions

We use multiple publicly available datasets, each under specific licenses. DailyDialog is licensed under CC BY-NC-SA 4.0 (Li et al., 2017). LCCC is released under the MIT License (Wang et al., 2022). PROSOCIALDIALOG and SODA are both licensed under CC-BY-4.0 (Kim et al., 2022, 2023). NORMDIAL (Li et al., 2023) is publicly released on GitHub, but no formal license is specified at the time of writing. All datasets are used in compliance with their respective licenses for research purposes.

We utilize several open-source language models under their respective licenses. LLaMA-3 models are released under the Llama 3 Community License Agreement (Dubey et al., 2024). Qwen-2.5 models are distributed under the Apache License 2.0 (Team, 2024). All models are used in accordance with their licensing terms for research purposes.

In data generation, we used a temperature of 0.7, while in evaluation, we used zero temperature.

E.2 Refinement Quality (RQ)

To evaluate the effectiveness of exemplar-based iterative refinement, we define three targeted evaluation dimensions: (1) **Norm Alignment**, assessing the degree to which the refined text conforms to the intended social norm; (2) **Language Quality**, measuring improvements in grammaticality, fluency, and stylistic appropriateness; and (3) **Semantic Fidelity**, evaluating whether the core meaning and intent of the original input are preserved post-refinement. These criteria are applied consistently across human and LLM-based evaluations. Prompt templates for each dimension are detailed below.

E.2.1 Norm Alignment Prompt.

Norm Alignment Prompt

```
Evaluation Instruction:
You are a domain expert who evaluates ONLY Norm Alignment—how well each text aligns with the given Social Norm. Ignore grammar or style.
Inputs:
Social Norm: {social_norm}
Initial Text: {initial}
Refined Text: {refined}
Scoring: Give each text an integer score 1-5 (1 = completely unrelated, 5 = perfectly aligned).
```

E.2.2 Language Quality Prompt.

Language Quality Prompt

```
Evaluation Instruction:
You are a professional copy-editor judging ONLY Language Quality—grammar, fluency, and naturalness. Ignore meaning preservation.
Inputs:
```

Category	English Subnorm Example	Korean Subnorms Example	Chinese Subnorm Example
Apology	In American culture, apologies are not typically accompanied by physical gestures like bowing, even when addressing older or higher-status individuals. What matters more is the sincerity conveyed through eye contact, tone, and word choice.	윗사람에게 사과할 때는 '변명 없이 바로 사과'하는 것이 중요하게 여겨진다. (When apologizing to a superior, it is considered important to apologize immediately without making excuses.)	In Chinese culture, bowing is not a common component of everyday apologies, even when addressing someone older or of higher status. Instead, sincerity is conveyed through respectful language, modest tone, and a deferential attitude.
Compliment	In the workplace, direct praise for effort is acceptable and often encouraged, especially from a supervisor to a subordinate.	한국 문화에서는 직장 내에서 상급자가 하급자의 노력을 인정할 때, 과한 칭찬보다는 "수고했어요" 또는 "고생 많았어요"와 같은 겸손한 표현이 더 적절하다고 여겨진다. ("In Korean culture, when a superior acknowledges the efforts of a subordinate in the workplace, modest expressions such as 'You've worked hard' or 'Thank you for your effort' are considered more appropriate than excessive praise.")	In Chinese workplace culture, when a superior acknowledges the effort of a subordinate, modest expressions such as "辛苦了" or "做得不错" are preferred, as excessive compliments can make the recipient feel awkward.
Condolence	When attending a funeral or memorial service, offering respectful and sincere condolences in a calm tone is considered basic etiquette.	한국 문화에서 장례식장에 방문했을 때는 고인의 명복을 빌며 유가족에게 조심스럽고 정중한 말투로 위로를 건네는 것이 기본적인 예의로 여겨진다. (In Korean culture, when visiting a funeral hall, it is considered basic etiquette to pray for the repose of the deceased and to offer words of comfort to the bereaved family in a cautious and respectful manner.)	In Chinese culture, it is customary to offer condolences in a calm and respectful tone while expressing wishes for the deceased's peace. "节哀顺变" is often used.
Criticism	In the workplace, offering constructive criticism in a direct but tactful way is generally acceptable.	직장 내에서 후배나 동료의 실수를 지적할 때는 직접적인 언급보다 부드럽게 제안하는 방식이 선호된다. (In the workplace, when pointing out a mistake made by a junior colleague or peer, a gentle and suggestive approach is preferred over direct criticism.)	In Chinese workplace culture, indirect language and soft suggestions are preferred to preserve face and avoid confrontation.
Empathy	In American culture, emotional validation is often prioritized over immediately offering solutions when someone is going through a hard time.	한국 문화에서는 누군가가 힘든 일을 겪었을 때 문제 해결보다 정서적 지지가 우선된다. (In Korean culture, when someone is going through a difficult time, emotional support takes precedence over problem-solving.)	When someone is going through a difficult time, it is generally more appropriate in Chinese culture to acknowledge their emotions first rather than rushing to provide solutions.
Greeting	When meeting a supervisor or someone in a higher position for the first time, a firm handshake, eye contact, and a polite greeting are considered standard etiquette.	한국 문화에서는 직장 상사를 처음 만났을 때 고개를 숙이며 "안녕하십니까"라고 인사하는 것이 예절이다. (In Korean culture, when meeting a workplace superior for the first time, it is considered proper etiquette to bow slightly and greet them by saying, "Annyeong hasimnikka" (a formal way of saying hello)	In Chinese workplace culture, it is customary to use a polite verbal greeting like "您好" while maintaining respectful posture and eye contact.
Leave-taking	At the end of a meeting or formal gathering, expressing appreciation for others' time and contributions is common.	한국 문화에서는 공식적인 만남 후 "수고 많으셨습니다" 라고 인사하며 헤어지는 것이 예의다. (In Korean culture, it is considered polite to part ways after a formal meeting by saying, "You've worked hard" (sugo manhasyeotsseumnida))	In Chinese culture, after a meeting or formal gathering, polite farewell phrases such as "辛苦了" or "谢谢大家的配合" are customary.
Persuasion	When persuading someone in a higher position, respectful directness is more common than indirect suggestions.	한국 문화에서는 윗사람을 설득할 때는 조심스러운 제안을 통해 접근하는 것이 일반적이다. (In Korean culture, when trying to persuade a superior, it is common to approach through cautious and respectful suggestions.)	In Chinese culture, indirect suggestions such as "您看是否可以考虑一下我的建议" are used to maintain politeness.
Request	When making a request to a superior, polite but direct phrasing is acceptable and even expected.	한국 문화에서는 윗사람에게 부탁할 때 "부탁드려도 될까요?"처럼 간접적이고 정중한 표현을 사용하는 것이 예의다. (In Korean culture, when making a request to a superior, it is considered polite to use indirect and respectful expressions such as, "May I ask a favor?")	In Chinese culture, it is customary to use indirect and respectful phrasing such as "不知道方不方便麻烦您一下？"
Respect	When offering opinions to someone in a higher position, respectful yet confident language is generally expected.	한국 문화에서는 "제 생각이 틀릴 수도 있지만…"처럼 조심스럽게 말문을 여는 것이 기본이다. (In Korean culture, it is customary to begin speaking cautiously with phrases such as, "I might be wrong, but...")	In Chinese culture, deferential phrases like "我有一点不成熟的想法" are used to show humility and respect.
Responding to Compliments	In American culture, it is generally considered appropriate to accept compliments graciously and confidently.	한국 문화에서는 칭찬을 받을 때 "아유, 그런 말씀 마세요"처럼 겸손하게 반응하는 것이 미덕이다. (In Korean culture, when receiving a compliment, it is considered virtuous to respond with humility, such as by saying, "Oh, please don't say that.")	In Chinese culture, people respond to compliments with humility using phrases like "哪里哪里" or "不敢当."
Thanks	In American workplaces, expressing thanks to a superior typically focuses on appreciation without explicitly referencing their rank.	한국 문화에서는 감사 표현 시 상대의 지위와 기여를 함께 언급하는 것이 공손한 방식으로 여겨진다. (In Korean culture, when expressing gratitude, it is considered polite to mention the other person's status and contribution together.)	In Chinese culture, gratitude often includes both the help and the role: "谢谢您，张经理，您帮我解决了问题。"

Figure 3: Subnorm examples

```
Initial Text: {initial}
Refined Text: {refined}
Scoring: Give each text an integer score 1–5 (1 = very poor language, 5 = native-level fluent).
```

E.2.3 Semantic Fidelity Prompt.

Semantic Fidelity Prompt

```
Evaluation Instruction:
You are a bilingual reviewer judging ONLY Semantic Fidelity—how faithfully the Refined text keeps the original meaning and intent of the Initial text. Ignore style and social-norm fit.
Inputs:
Initial Text: {initial}
Refined Text: {refined}
Scoring: Give each text an integer score 1–5 (1 = meaning lost / contradictory, 5 = meaning identical).
```

E.3 Dialogue Quality (DQ)

To comprehensively evaluate dialogue quality, we assess multi-turn conversations across six dimensions that capture linguistic fluency, pragmatic coherence, and contextual relevance. The evaluation criteria are as follows: (1) Consistency, which measures logical coherence across dialogue turns; (2) Naturalness, which assesses fluency and idiomaticity of language; (3) Relevance, which evaluates how well the dialogue reflects the given scenario and situation; (4) Emotional Appropriateness, which measures the suitability of tone and affective expression; (5) Social Norm Appropriateness, which determines the degree of alignment with the intended norm; and (6) Scenario Coherence, which assesses semantic continuity between the scenario-situation and the dialogue. Each dimension is evaluated using dedicated prompt templates applied to both human annotators and LLM-based evaluators. The full set of prompt formulations is provided in Table 23, 24, 25, 26, 27, and 28.

E.4 Generalization Quality (GQ)

To assess model generalization beyond the training distribution, we evaluate whether models fine-tuned on our dataset can produce socially appro-

priate and contextually coherent responses in out-of-domain scenarios. Specifically, we use dialogue contexts sampled from two external benchmarks: DailyDialog (for English) and LCCC (for Chinese). Each model is prompted to generate five-turn continuations for these contexts. The generated outputs are then evaluated through A/B preference tests conducted by both human annotators and LLM-based evaluators. Judgments are based on criteria such as naturalness and alignment with social norms. To ensure robustness and comparability, the evaluation spans multiple language model architectures. Prompt templates for continuation generation and preference elicitation are provided in Table 29.

F Comparative Analysis of Generation in Low-Resource Settings

F.1 Examples of Common Errors

To better understand the limitations of prompt-only generation in low-resource languages, we present qualitative examples of common errors observed in dialogues from typologically diverse languages. These examples underscore the challenges that arise when language models are required to generate pragmatically appropriate outputs in socioculturally complex settings without sufficient grounding.

Figure 16 illustrates two prevalent generation errors: (1) lexical redundancy, where intensifiers or formulaic actions such as “lowered his head and said” are repeated across multiple turns, and (2) tone mismatches, where informal phrasing appears in contexts that require formal or respectful speech levels. Both issues indicate a lack of contextual awareness regarding interpersonal roles and emotional nuance, which are crucial in conversation norms for languages with rich honorific systems.

Figure 17 presents analogous issues. We observe (1) emotional redundancy, with exaggerated repetitions of intensifiers (e.g., "really really very very") that undermine sincerity; (2) honorific inconsistency, where formal pronouns are combined with casual address terms within the same sentence; and (3) register mixing, where archaic written forms are abruptly followed by colloquial speech. These inconsistencies reflect the model’s difficulty in maintaining coherent style and role-appropriate politeness strategies.

Collectively, these examples provide qualitative evidence for the sociolinguistic brittleness of prompt-only methods in languages with morphosyntactic complexity and hierarchical speech conventions. They further motivate our refinement-based framework, which explicitly encodes cultural subnorms and stylistic expectations to produce fluent, contextually aligned, and socially coherent outputs in low-resource settings.

F.2 Additional Pilot Refinement Experiment

We recognize that languages with divergent pragmatic and honorific systems (e.g., Arabic, Swahili, Hindi) pose additional challenges. In such cases, our exemplar-based refinement component is crucial for the early injection of sociocultural constraints, especially in the absence of large annotated corpora. To test generalizability beyond East Asian typologies, we conducted pilot refinement in Malay and Urdu, two pragmatically distinct languages. Using the same evaluation setup as in Table 4, we find the results align closely with those from our main experiments, shown in Table 22. These findings demonstrate the framework’s generalizability to typologically and culturally distinct languages and will be included in the camera-ready version as initial evidence of broader applicability. We plan to extend to additional low-resource languages, including Arabic and Swahili.

Input: Languages $\mathcal{L}$, Norm Categories $\mathcal{C}$, Dialogue Types $\mathcal{T}$
Output: Annotated Dialogues $\mathcal{D}$
foreach *language* l *in* $\mathcal{L}$ **do**
  foreach *category* c *in* $\mathcal{C}$ **do**
    Generate subnorm set $\mathcal{N}_{l,c}$;
  end
end
foreach *language* l *in* $\mathcal{L}$ **do**
  foreach *category* c *in* $\mathcal{C}$ **do**
    foreach *subnorm* n *in* $\mathcal{N}_{l,c}$ **do**
      foreach *dialogue type* t *in* $\mathcal{T}$ **do**
        Generate Scenario $S_{l,n,t}$;
        Generate Situation $T_{l,n,t}$ based on $S_{l,n,t}$;
      end
    end
  end
end
foreach *language* l *in* $\mathcal{L}$ **do**
  foreach *category* c *in* $\mathcal{C}$ **do**
    foreach *subnorm* n *in* $\mathcal{N}_{l,c}$ **do**
      Expert manually refines exemplar pair $E_{l,n}$;
      repeat
        Refine Scenario-Situation pairs $(S,T)_{l,n,*}$ aligned to $E_{l,n}$;
        Evaluate refinement quality $Q_{l,n}$;
      until *Refinement Quality* $Q_{l,n}$ *meets threshold*;
    end
  end
end
foreach *refined pair* $(S,T)_{l,n,t}$ **do**
  Generate Dialogue $D_{l,n,t}$ using (c,n,S,T) as input;
  Annotate each turn in $D_{l,n,t}$ with norm and reaction labels;
end
return *All annotated dialogues* $\mathcal{D}$

Algorithm 1: Multicultural social norm dialogue Generation Pipeline

Language	Condition	RQ (GPT-4o) Avg.	RQ (Human Evaluation) Avg.
Malay	Initial	3.623	2.666
	Refined	**4.830**	**4.333**
Urdu	Initial	3.217	3.166
	Refined	**4.777**	**4.222**

Table 22: Refinement quality (RQ) evaluation in Malay and Urdu. Scores are based on a 5-point Likert scale, averaged over all evaluation dimensions and aggregated across both LLM-based and human raters.

Category	Subnorm
Apology	Bowing to elders when apologizing is not considered necessary or meaningful. In American culture, apologies are not typically accompanied by physical gestures like bowing, even when addressing older or higher-status individuals. What matters more is the sincerity conveyed through eye contact, tone, and word choice. For example, one might say, "I'm really sorry, sir. I take full responsibility," while maintaining a calm and respectful demeanor.

Scenario(After Refinement)

In a high school classroom, Jake accidentally bumps into his history teacher,
Mrs. Thompson, and immediately says, "I'm so sorry, Mrs. Thompson. I didn't mean to run into you," while looking her in the eye.

Situation(After Refinement)

As Jake hurried down the crowded hallway, balancing a stack of books, he accidentally collided with Mrs. Thompson just as she stepped out of her classroom. The books tumbled to the floor, and Jake's cheeks flushed with embarrassment. Quickly regaining his composure, he looked Mrs. Thompson in the eye and said, "I'm so sorry, Mrs. Thompson. I didn't mean to run into you." Mrs. Thompson smiled kindly, brushing off her sleeve, and replied, "It's alright, Jake. Just be a little more careful next time." Jake nodded, grateful for her understanding, and stooped to gather his books.

Dialogue (After Refinement)

Jake: Oh no—Mrs. Thompson, I'm so sorry! I didn't see you there.
Mrs. Thompson: That's alright, Jake. Are you okay?
Jake: Yeah, I'm fine. I really didn't mean to bump into you. I should have been paying more attention.
Mrs. Thompson: Don't worry about it. The hallways are crowded today. Just try to be a bit more careful next time.
Jake: I will. Thanks for understanding. Let me get these books picked up.
Mrs. Thompson: Thank you, Jake. Let me help you with those.
Jake: Thanks, Mrs. Thompson. I really appreciate it.
Mrs. Thompson: No problem at all. Have a good rest of your day, and watch out for hallway traffic!
Jake: I will! You too, Mrs. Thompson.

Figure 4: Adherence norm example (EN)

Consistency Prompt

Section	Content
Evaluation Instruction	You are a professional dataset auditor for social-norm dialogues. You are given a culture category. Your task is to evaluate only the *Consistency* of the dialogue. Ignore grammar, fluency, or style. Focus only on whether the dialogue is logically and contextually consistent throughout.
Parameter	`culture`, `norm`, `dialogue`
Evaluation Question	Assuming the dialogue adheres to the given social norm, are all utterances logically and emotionally coherent with one another? – Do characters maintain a consistent attitude, tone, and perspective throughout? – Are there any contradictions or abrupt shifts in reasoning, emotion, or information? – Does the dialogue flow smoothly without unexpected or unjustified changes?
Scoring Criteria	1 = Major inconsistencies or contradictions 3 = Somewhat inconsistent or awkward transitions 5 = Fully consistent and coherent throughout

Table 23: Evaluation prompt structure for *Consistency*.

Category	Subnorm
Apology	윗사람에게 사과할 때는 '변명 없이 바로 사과'하는 것이 중요하게 여겨진다. (When apologizing to a superior, it is considered important to apologize directly without making excuses.)
Scenario(Before Refinement)	**Scenario(After Refinement)**
민수는 회의 중에 잘못된 정보의 내용을 발표하고 난 후 잘못 전달된 내용인걸 파악하여, 사장님에게 변명 없이 "죄송합니다"라고 사과한다. (Minsu, after presenting incorrect information during a meeting and realizing the mistake, apologized to the president without making excuses by saying, "I'm sorry.")	사원인 민수는 회의 중 잘못된 정보를 발표한 뒤, 실수를 바로 인지하고 사장님께 변명 없이 "죄송합니다"라고 정중히 사과한다. (As an employee, Minsu recognized his error immediately after presenting the wrong information in a meeting and offered a polite apology to the president without excuses, saying, "I'm sorry.")
Situation(Before Refinement)	**Situation(After Refinement)**
회의실에서 민수는 발표를 하던 중 자신이 잘못된 숫자를 제시한 것을 깨달았다. 발표가 끝난 후, 사장님이 곧바로 그 부분에 대해 질문하자 민수는 얼굴이 붉어졌다. 그는 가벼운 떨림을 느끼며, "죄송합니다 사장님, 프레젠테이션을 진행할 때 잘못된 정보 드렸습니다"라고 고개를 숙이며 말했다. 사장님은 잠시 침묵하다가, "그래, 다음엔 자료 준비를 더 주의해서 프레젠테이션을 진행해줘 "라고 부드럽게 대답했다. (In the meeting room, while giving his presentation, Minsu realized that he had presented incorrect numbers. After the presentation ended, the president immediately asked about that part, and Minsu's face turned red. Feeling a slight tremor, he bowed his head and said, "I'm sorry, President. I gave the wrong information during the presentation." The president paused for a moment before replying gently, "Alright, please be more careful in preparing your materials for the next presentation.")	회의 중 발표를 진행하던 민수는 도중에 자신이 잘못된 수치를 제시했다는 사실을 깨달았다. 발표를 마친 직후, 사장님이 조용히 그 부분을 짚어주시자 민수의 얼굴이 붉게 달아올랐다. 잠시 말을 잃은 그는 떨리는 목소리로 고개를 숙이며 말했다. "죄송합니다, 사장님. 자료를 충분히 검토하지 못하고 잘못된 정보를 드렸습니다." 사장님은 한동안 민수를 바라보시다가 차분한 목소리로 말씀하셨다. "괜찮아요. 누구나 실수할 수 있으니, 다음엔 더 꼼꼼히 준비해 주세요." 회의실은 다시 조용하고 차분한 분위기로 돌아갔고, 민수는 고개를 깊이 숙이며 감사의 뜻을 전했다. (During his presentation, Minsu realized that he had presented an incorrect figure. Immediately after finishing, the president quietly pointed out the mistake, and Minsu's face flushed red. Momentarily at a loss for words, he bowed his head and spoke in a trembling voice, "I'm sorry, President. I failed to review the data thoroughly and gave the wrong information." The president looked at him for a while before speaking in a calm tone, "It's alright. Anyone can make mistakes, but please prepare more carefully next time." The meeting room soon returned to a quiet, composed atmosphere, and Minsu lowered his head deeply to express his gratitude.)
	Dialogue (After Refinement)
민수: 죄송합니다, 사장님. 발표 중에 잘못된 정보를 드렸습니다. **사장님:** 그래, 민수 씨. 실수한 부분을 바로 인정해줘서 고마워요. **민수:** 앞으로는 더 꼼꼼하게 자료를 준비하겠습니다. **사장님:** 좋아요. 다음에는 조금 더 주의해서 발표해 주면 좋겠어요. **민수:** 네, 꼭 그렇게 하겠습니다. 죄송합니다. **(Minsu**: I'm sorry, President. I gave incorrect information during the presentation. **President**: That's alright, Minsu. Thank you for admitting the mistake right away. **Minsu**: I'll make sure to prepare the materials more thoroughly from now on. **President**: Good. Next time, I'd appreciate it if you could be a bit more careful in your presentation. **Minsu**: Yes, I will. I'm truly sorry.)	**민수:** 사장님, 정말 죄송합니다. 자료를 제대로 확인하지 못해 잘못된 정보를 드렸습니다. **사장님:** 음... 괜찮아요, 민수 씨. 준비하다 보면 그런 실수도 있을 수 있어요. **민수:** 다시는 이런 일이 없도록 더 꼼꼼히 준비하겠습니다. **사장님:** 네, 다음에는 더 신경 써서 발표해 주세요. **민수:** 네, 유의하겠습니다. 죄송합니다. **사장님:** 그래도 이렇게 모두가 있는 자리에서 책임 있게 사과해 주셔서 좋았습니다. **민수:** 감사합니다, 사장님. 앞으로는 발표 전에 두세 번 더 꼼꼼히 확인하겠습니다. **사장님:** 좋아요. 이번 경험이 민수 씨에게도 좋은 계기가 되길 바랍니다. **민수:** 네, 그렇게 하겠습니다. 다시 한 번 기회 주셔서 감사합니다. **Minsu**: President, I'm truly sorry. I didn't check the data thoroughly and ended up giving you the wrong information. **President**: Hmm... it's alright, Minsu. Mistakes like that can happen while preparing. **Minsu**: I'll make sure to prepare more carefully so this never happens again. **President**: Yes, please pay more attention in your next presentation. **Minsu**: Yes, I will keep that in mind. I'm sorry. **President**: Still, I appreciate that you took responsibility and apologized in front of everyone. **Minsu:** Thank you, President. From now on, I'll review everything two or three times more carefully before presenting. **President**: Good. I hope this experience becomes a meaningful lesson for you as well. **Minsu**: Yes, it will. Thank you once again for giving me another chance.

Figure 5: Adherence norm example (KR)

Category	Subnorm
Apology	When apologizing to someone in a higher position, offering excuses beforehand is typically seen as inappropriate in Chinese culture. A direct assumption of responsibility is more acceptable and considered polite. For instance, saying “这次是我的责任，没有任何借口” immediately shows sincerity and self-awareness without appearing defensive or disrespectful.
Scenario(Before Refinement)	**Scenario(After Refinement)**
During the meeting, Li Ming said: “I was late, this is my fault, I don’t have excuse.”	During a team meeting, Li Ming apologized directly to 王经理 by saying, “王经理，这次迟到是我的责任，没有任何借口.”
Situation(Before Refinement)	**Situation(After Refinement)**
Monday morning, Li Ming came late. He said, 'Manager Wang, this is my responsibility for the late.' He didn’t bow. Manager Wang nodded. The meeting continued. People thought it was weird.	On Monday morning, Li Ming entered the meeting room late. Without excuse, he gave a slight 鞠躬 and said, '王经理，这次迟到是我的责任，没有任何借口.' 王经理 nodded. The meeting continued smoothly. Colleagues noted his sincerity and professionalism.
Dialogue (Before Refinement)	**Dialogue (After Refinement)**
李明: 王经理，对不起，今天我迟到了，这是我的责任，没有任何借口。 **王经理:** 好，坐下吧，会议继续。 **李明:** 谢谢王经理，下次我会注意。 **王经理:** 行，下不为例。 **Li Ming**: Manager Wang, I’m sorry. I was late today. This is my responsibility, and I have no excuses. **Manager Wang**: Alright, have a seat. Let’s continue the meeting. **Li Ming**: Thank you, Manager Wang. I’ll be more careful next time. **Manager Wang**: Fine, but don’t let it happen again.	**李明:** (急匆匆进门，微微鞠躬) 王经理，对不起，我来晚了。是我时间安排不当，责任在我，没有任何借口。 **王经理:** 好，知道就行。下不为例，赶紧坐下吧，我们继续会议。 **李明:** 是，王经理。给大家添麻烦了，真的很抱歉。 **王经理:** 没关系，注意就好。大家互相体谅，共同进步。 **同事A:** 没事李明，我们也都有过类似情况。 **李明:** 谢谢大家的理解，以后我一定提前安排好，确保不再迟到。 **王经理:** 好，那我们接着往下讲。 **Li Ming**: (hurrying into the room, bowing slightly) Manager Wang, I’m sorry I’m late. I didn’t manage my time properly. The responsibility is mine, and I have no excuses. **Manager Wang**: Alright, as long as you understand. Don’t let it happen again. Take a seat, let’s continue the meeting. **Li Ming**: Yes, Manager Wang. I’m truly sorry for causing trouble for everyone. **Manager Wang**: It’s fine, just pay attention in the future. We should all be understanding and make progress together. **Colleague A**: It’s okay, Li Ming. We’ve all had similar situations before. **Li Ming**: Thank you for your understanding, everyone. From now on, I’ll plan ahead to make sure I’m never late again. **Manager Wang**: Good, then let’s continue.

Figure 6: Adherence norm example (ZH)

Turn-Level Annotated Dialogue – Adhered (EN)	Turn-Level Annotated Dialogue – Adhered (KR)	Turn-Level Annotated Dialogue – Adhered (ZH)
Jake: Oh no—Mrs. Thompson, I'm so sorry! I didn't see you there. **(Adhered | APO | Jake expresses a sincere apology for his mistake, following the social norm.)** **Mrs. Thompson:** That's alright, Jake. Are you okay? **(Adhered | EMP | Mrs. Thompson shows emotional concern for Jake's well-being, expressing empathy.)** **Jake:** Yeah, I'm fine. I really didn't mean to bump into you. I should have been paying more attention. **(Adhered | JUS | Jake explains his intention and admits his mistake, offering a justification.)** **Mrs. Thompson:** Don't worry about it. The hallways are crowded today. Just try to be a bit more careful next time. **(Adhered | SUG | Mrs. Thompson gives gentle advice while normalizing the situation.)** **Jake:** I will. Thanks for understanding. Let me get these books picked up. **(Adhered | THX | Jake thanks Mrs. Thompson for her understanding and takes action to correct the situation.)** **Mrs. Thompson:** Thank you, Jake. Let me help you with those. **(Adhered | THX | Mrs. Thompson expresses gratitude and offers support.)** **Jake:** Thanks, Mrs. Thompson. I really appreciate it. **(Adhered | THX | Jake repeats his appreciation for the help offered.)** **Mrs. Thompson:** No problem at all. Have a good rest of your day, and watch out for hallway traffic! **(Adhered | SUG | Mrs. Thompson offers light-hearted advice to conclude the interaction warmly.)** **Jake:** I will! You too, Mrs. Thompson. **(Adhered | ACK | Jake acknowledges her message and reciprocates the well wishes.)**	**민수:** 사장님, 정말 죄송합니다. 자료를 제대로 확인하지 못해 잘못된 정보를 드렸습니다. **(Adhered | APO | 민수가 자신의 실수에 대해 사과하며 책임을 인정하는 발언으로, 사회적 규범을 잘 따르고 있음.)** **Minsu**: President, I'm truly sorry. I failed to check the data thoroughly and gave you the wrong information. (Adhered | APO | Minsu apologizes for his mistake and takes responsibility, showing adherence to social norms.) **사장님:** 음... 괜찮아요, 민수 씨. 준비하다 보면 그런 실수도 있을 수 있어요. **(Adhered | EMP | 사장이 실수를 이해하고 공감하며 감정을 완화시키는 표현을 사용함.)** **President**: Hmm… it's alright, Minsu. Mistakes like that can happen while preparing. (Adhered | EMP | The president shows understanding and empathy, using words that ease the emotional tension.) **민수:** 다시는 이런 일이 없도록 더 꼼꼼히 준비하겠습니다. **(Adhered | SUG | 민수가 스스로 개선 방안을 제시하며 책임감을 표현함.)** **Minsu**: I'll make sure to prepare more carefully so this never happens again. (Adhered | SUG | Minsu proposes his own improvement plan, expressing responsibility.) **사장님:** 네, 다음에는 더 신경 써서 발표해 주세요. **(Adhered | SUG | 사장이 조언과 당부를 통해 향후 주의할 점을 전달함.)** **President**: Yes, please be more attentive in your next presentation. (Adhered | SUG | The president gives advice and a reminder about what to be careful of in the future.) **민수:** 네, 유의하겠습니다. 죄송합니다. **(Adhered | APO | 민수가 사장의 조언을 수용하고 다시 한 번 사과함.)** **Minsu**: Yes, I'll keep that in mind. I'm sorry. (Adhered | APO | Minsu accepts the president's advice and apologizes once again.) **사장님:** 그래도 이렇게 모두가 있는 자리에서 책임 있게 사과해 주셔서 좋았습니다. **(Adhered | THX | 사장이 민수의 책임감 있는 태도를 긍정적으로 평가하며 감사를 표현함.)** **President**: Still, I appreciate that you took responsibility and apologized here in front of everyone. (Adhered | THX | The president positively evaluates Minsu's responsible attitude and expresses appreciation.) **민수:** 감사합니다, 사장님. 앞으로는 발표 전에 두세 번 더 꼼꼼히 확인하겠습니다. **(Adhered | THX | 민수가 감사를 표현하고, 재발 방지를 위한 구체적 노력을 언급함.)** **Minsu**: Thank you, President. From now on, I'll review everything two or three times more carefully before presenting. (Adhered | THX | Minsu expresses gratitude and mentions specific efforts to prevent recurrence.) **사장님:** 좋아요. 이번 경험이 민수 씨에게도 좋은 계기가 되길 바랍니다. **(Adhered | EMP | 사장이 격려와 응원의 메시지를 전하며 긍정적인 마무리를 지음.)** **President**: Good. I hope this experience becomes a valuable lesson for you as well. (Adhered | EMP | The president closes on a positive note with encouragement and support.) **민수:** 네, 그렇게 하겠습니다. 다시 한 번 기회 주셔서 감사합니다. **(Adhered | THX | 민수가 다짐과 함께 감사의 표현으로 대화를 마무리함.)** **Minsu**: Yes, I will. Thank you again for giving me another chance. (Adhered | THX | Minsu ends the conversation with determination and an expression of gratitude.)	**李明:** （急匆匆进门，微微鞠躬）王经理，对不起，我来晚了。是我时间安排不当，责任在我，没有任何借口。 **(Adhered | APO | Li Ming delivers a sincere and formal apology, takes full responsibility, and avoids making excuses, aligning with Chinese workplace norms.)** **Li Ming**: (hurrying into the room, bowing slightly) Manager Wang, I'm sorry I'm late. I didn't manage my time properly. The responsibility is mine, and I have no excuses. **王经理:** 好，知道就行。下不为例，赶紧坐下吧，我们继续会议 **(Adhered | EMP | Manager Wang accepts the apology in a restrained and directive tone, maintaining authority while acknowledging the apology.)** **Manager Wang**: Alright, as long as you understand. Don't let it happen again. Sit down quickly, let's continue the meeting. **李明:** 是，王经理。给大家添麻烦了，真的很抱歉。 **(Adhered | APO | Li Ming further apologizes to the group, reinforcing accountability and politeness.)** **Li Ming**: Yes, Manager Wang. I'm truly sorry for causing trouble for everyone. **王经理:** 没关系，注意就好。大家互相体谅，共同进步。 **(Adhered | EMP | Manager Wang offers emotional support and promotes mutual understanding, consistent with collectivist values.)** **Manager Wang**: It's fine, just be mindful from now on. Let's all be understanding and make progress together. **同事A:** 没事李明，我们也都有过类似情况。 **(Adhered | EMP | The colleague expresses empathy and normalizes the mistake, which is socially supportive behavior.)** **Colleague A**: It's okay, Li Ming. We've all had similar situations before. **李明:** 谢谢大家的理解，以后我一定提前安排好，确保不再迟到 **(Adhered | THX | Li Ming expresses gratitude and a commitment to improvement, showing both humility and responsibility.)** **Li Ming**: Thank you all for your understanding. From now on, I'll plan ahead to make sure I'm never late again. **王经理:** 好，那我们接着往下讲。 **(Not relevant | N/A | This utterance shifts the focus back to the meeting agenda and is not directly related to the evaluation of social norm adherence. It serves a purely organizational function.)** **Manager Wang**: Good, then let's move on.

Figure 7: Turn-level annotated dialogue example – Adherence

Category	Subnorm
Apology	Bowing to elders when apologizing is not considered necessary or meaningful. In American culture, apologies are not typically accompanied by physical gestures like bowing, even when addressing older or higher-status individuals. What matters more is the sincerity conveyed through eye contact, tone, and word choice. For example, one might say, "I'm really sorry, sir. I take full responsibility," while maintaining a calm and respectful demeanor.

Scenario(After Refinement)

In a college study group, Alex accidentally interrupts their professor, Dr. Patel, during a discussion, and their initial response comes off as dismissive rather than apologetic.

Situation(After Refinement)

During a lively seminar, Alex, eager to share a point, jumps in and cuts off Dr. Patel mid-sentence, saying, "Yeah, but I think it's more about social factors," without looking up from their notes. The room falls quiet for a moment. Dr. Patel raises her eyebrows slightly, caught off guard. Realizing the interruption and how dismissive it sounded, Alex pauses, then looks up to meet her eyes. "I'm sorry, Dr. Patel—I didn't mean to interrupt. Please go ahead," they say sincerely. Dr. Patel's expression softens, and she gives a small nod before continuing. The tension in the room eases, and Alex listens more attentively, relieved to have handled the moment with maturity.

Dialogue (After Refinement)

Alex: (cutting in, eyes still on notebook) Yeah, but I just think social factors are actually more relevant—
Dr. Patel: (pauses, lifting an eyebrow, clearly caught off guard)
Alex: (keeps writing, unaware of the silence)
Dr. Patel: (calm but firm) Alex, hold on—I hadn't finished my point.
Alex: (looks up, noticing the shift in tone) Oh—sorry, Dr. Patel. I didn't mean to jump in like that.
Dr. Patel: (slightly frowning) I know you're eager to contribute, and that's great. But let's make sure we're giving everyone space to finish their thoughts.
Alex: (face flushing, now fully engaged) You're right. I apologize—I got ahead of myself. Please, go ahead.
Dr. Patel: (nodding, tone softens) Thanks, Alex. I appreciate that. We're all here to learn from each other, and that works best when we listen, too.
Alex: (genuinely) Absolutely. I'll be more mindful going forward. Really sorry again.
Dr. Patel: (smiling) No worries. Let's pick up where we left off...
Alex: (sits up straighter, fully focused)

Figure 8: Violation-to-resolution(V2R) norm example (EN)

Naturalness Prompt

Section	Content
Evaluation Instruction	You are a professional dataset auditor for social-norm dialogues. Your task is to evaluate only the *Naturalness* of the dialogue. Ignore whether the response is factually correct or norm-appropriate. Focus on how naturally the dialogue would sound to a native speaker.
Parameter	`dialogue`
Evaluation Question	Does the dialogue sound natural and fluent as if spoken by native speakers in a real-world situation? – Are the expressions, tone, and word choices contextually appropriate and idiomatic? – Do the conversational turns flow smoothly without sounding robotic or overly scripted? – Are there any awkward phrases or unnatural sentence structures?
Scoring Criteria	1 = Extremely unnatural or robotic 3 = Somewhat awkward or artificial 5 = Very natural, fluent, and human-like

Table 24: Evaluation prompt structure for *Naturalness*.

Category	Subnorm
Apology	윗사람에게 사과할 때는 '변명 없이 바로 사과'하는 것이 중요하게 여겨진다.

Scenario(Before Refinement)	Scenario(After Refinement)
민수는 회의 시간에 잘못된 정보를 말했다. 사장님이 "이거 숫자가 좀 이상한데요?"라고 하시자, 민수는 갑자기 "아, 그게... 자료가 좀 헷갈리게 와가지고요. 원래 그런 건 아닌데..."라며 말끝을 흐렸다. 사장님의 얼굴이 딱 굳어졌고, 회의실이 이상하게 조용해졌다. 민수는 주변을 두리번거리다가 갑자기 말을 꺼냈다. "그게 전부 저의 불찰입니다. 죄송합니다, 진짜요... 제가 잘못 본 것 같아요." (During the meeting, Minsu gave incorrect information. When the president said, "These numbers look a bit unusual," Minsu suddenly replied, "Ah, well... the data was sent in a confusing way. It wasn't supposed to be like that..." trailing off at the end. The president's face stiffened, and the meeting room fell into an awkward silence. Looking around uneasily, Minsu then spoke up: "It's entirely my fault. I'm sorry, truly... I must have misread it.")	사원 민수는 회의 중 중요한 수치를 잘못 발표했고, 사장님의 지적에 처음에는 "최근에 받은 자료가 혼동을 줄 수 있게 정리되어 있어서 그렇다"며 책임을 회피했다. 그러나 사장님의 굳은 표정과 무거운 회의 분위기를 느낀 민수는 곧 자신의 태도가 적절하지 않았음을 깨닫고, 사과의 말을 꺼낸다. (Employee Minsu presented an important figure incorrectly during a meeting, and when the president pointed it out, he initially tried to evade responsibility by saying, "The data I received recently was organized in a confusing way." However, sensing the president's stern expression and the heavy atmosphere in the meeting room, Minsu soon realized that his response had been inappropriate and offered an apology.)
Situation(Before Refinement)	**Situation(After Refinement)**
프로젝트 성과 발표 중 민수는 숫자를 잘못 말했다. 발표가 끝나고 사장님이 조용히, "민수 씨, 방금 제시한 수치가 기존이랑 다른데요?"라고 하셨다. 민수는 살짝 당황한 표정으로, "아... 그게요, 회계팀 자료가 좀 복잡하게 돼 있어서요. 뭔가 좀 이상하긴 했는데... 그냥 넘겼어요."라고 말하며 웃었다. 사장님은 아무 말 없이 민수를 뚫어지게 보셨고, 회의실 분위기는 뻣뻣하게 굳었다. 민수는 어색한 정적 속에서 어깨를 한번 으쓱하더니 작게 말했다. "저의 불찰입니다. 자료를 대충 본 건 아닌데, 좀 그렇게 됐네요. 죄송합니다, 사장님." (During the project performance presentation, Minsu reported an incorrect number. After the presentation ended, the president quietly said, "Minsu, the figure you just presented doesn't match the previous one." Looking slightly flustered, Minsu replied, "Ah... well, the accounting team's data was a bit complicated. I did notice something seemed off... but I just let it pass," and gave a small laugh. The president said nothing, staring intently at Minsu, and the atmosphere in the meeting room grew stiff. In the awkward silence, Minsu shrugged his shoulders slightly and then said quietly, "It's my oversight. It's not that I didn't review the data, but it turned out this way. I'm sorry, President.")	회의 중 사원 민수는 중요한 수치를 잘못 발표했고, 사장님의 지적에 "회계팀 자료가 중간에 바뀌어서..."라며 책임을 회피했다. 사장님의 표정이 굳고, 회의실이 잠시 정적에 잠기자 민수는 자신의 태도를 반성하며 말했다. "죄송합니다, 사장님. 변명부터 드린 점 깊이 사과드립니다. 제 불찰입니다." 사장님은 부드럽게, "그래요. 다음엔 더 꼼꼼히 확인해 주세요"라고 답했다. (During a meeting, employee Minsu presented an important figure incorrectly, and when the president pointed it out, he tried to evade responsibility by saying, "The accounting team's data was changed midway..." When the president's expression hardened and the room fell into a brief silence, Minsu reflected on his attitude and said, "I'm sorry, President. I sincerely apologize for making excuses first. It was my oversight." The president replied gently, "Alright. Please check more carefully next time)
Dialogue (Before Refinement)	**Dialogue (After Refinement)**
민수: (발표가 끝난 후) 아... 그게요, 회계팀 자료가 좀 복잡하게 돼 있어서요. 뭔가 좀 이상하긴 했는데... 그냥 넘겼어요. (어색하게 웃는다) **사장님:** (조용히 민수를 바라본다) 민수 씨, 중요한 숫자는 반드시 확인해주셔야 합니다. **민수:** (어깨를 으쓱하며 작게) 저의 불찰입니다. 자료를 대충 본 건 아닌데, 좀 그렇게 됐네요. 죄송합니다, 사장님. **사장님:** (잠시 침묵 후) 네, 앞으로는 이런 일이 없도록 주의해주시기 바랍니다. **민수:** (잠시 생각하다가) 네... 사실, 방금 제가 충분히 확인하지 못하고, 제대로 사과도 드리지 못한 것 같습니다. 괜히 변명을 먼저 드린 점 정말 죄송합니다. 전적으로 제 책임입니다. 앞으로는 이런 실수 반복하지 않겠습니다. **사장님:** (표정이 조금 누그러진다) 네, 민수 씨. 이렇게 바로 인정하고 사과해주셔서 고맙습니다. 누구나 실수할 수 있지만, 책임을 인정하고 빨리 바로잡는 게 더 중요합니다. **민수:** 네, 명심하겠습니다. 앞으로는 더 신중하게 준비하겠습니다. 다시 한 번 죄송합니다. **사장님:** 네, 좋습니다. 계속 힘내주세요, 민수 씨. **Minsu**: (after the presentation) Ah... well, the accounting team's data was a bit complicated. I did notice something seemed off... but I just let it pass. (smiles awkwardly) **President**: (quietly looking at Minsu) Minsu, you must make sure to verify important numbers. **Minsu**: (shrugging slightly, speaking softly) It's my oversight. It's not that I didn't review the data, but it turned out this way. I'm sorry, President. **President**: (after a brief pause) Alright, please be careful so this doesn't happen again. **Minsu**: (after thinking for a moment) Yes... in fact, I didn't check thoroughly just now, and I also didn't apologize properly. I'm truly sorry for making excuses first. It's entirely my responsibility. I won't repeat this mistake in the future. **President**: (expression softening a little) Yes, Minsu. Thank you for admitting and apologizing right away. Everyone can make mistakes, but acknowledging responsibility and correcting it quickly is more important. **Minsu**: Yes, I'll keep that in mind. I'll be more careful in my preparation from now on. I sincerely apologize once again. **President**: Good. Keep up the effort, Minsu.	**사장님:** 민수 씨, 이번 매출 보고서에서 지난달 수치가 잘못 기재된 것 같네요. 확인해 보셨어요? **민수:** (급하게 모니터를 보며) 아, 네... 아마 회계팀 자료가 중간에 바뀌어서 그럴 거예요. 저도 헷갈려서... **사장님:** (표정이 딱딱해지며) 다른 팀 때문이라는 건가요? (잠깐 정적, 민수는 사장님의 표정을 보고 잠시 생각한다) **민수:** 아... 죄송합니다, 사장님. 변명을 의도하려한 건 아닌데, 잘못을 제대로 인정하지 않고 말씀드린 점, 정말 죄송합니다. 제 실수입니다. **사장님:** (조금 부드러워진 목소리로) 그래요. 누구나 실수할 수는 있어요. 다음엔 좀 더 꼼꼼히 확인해 주세요. 지금 수정은 가능합니까? **민수:** 네, 바로 수정해서 보내드리겠습니다. 다시 한 번 죄송합니다. **사장님:** 앞으로는 문제가 생기면 바로 인정하고, 해결책도 같이 고민하도록 해요. 그게 더 신뢰를 줄 수 있어요. **민수:** 네, 명심하겠습니다. 감사합니다. **President**: Minsu, it looks like last month's figures were entered incorrectly in this sales report. Have you checked it? **Minsu**: (quickly looking at the monitor) Ah, yes... It's probably because the accounting team changed the data midway. I got confused too... **President**: (expression hardening) Are you saying it was the other team's fault? *(a brief silence, Minsu notices the president's expression and pauses to think)* **Minsu**: Ah... I'm sorry, President. I didn't mean to make excuses, but I realize I didn't properly acknowledge the mistake when I spoke. I sincerely apologize. This was my error. **President**: (voice softening a little) Alright. Anyone can make mistakes. Just be more thorough next time. Can it be corrected now? **Minsu**: Yes, I'll revise it right away and send it to you. I apologize once again. **President**: From now on, if a problem arises, admit it immediately and let's work on a solution together. That builds greater trust. **Minsu**: Yes, I'll keep that in mind. Thank you.

Figure 9: Violation-to-resolution(V2R) norm example (KR)

Category	Subnorm
Apology	When apologizing to someone in a higher position, offering excuses beforehand is typically seen as inappropriate in Chinese culture. A direct assumption of responsibility is more acceptable and considered polite. For instance, saying "这次是我的责任，没有任何借口" immediately shows sincerity and self-awareness without appearing defensive or disrespectful.

Scenario(Before Refinement)	Scenario(After Refinement)
Zhang Qiang was late. He said traffic and rain made him late.	Zhang Qiang arrives late to a meeting with 陈主任. Flustered, he gives excuses.
Situation(Before Refinement)	**Situation(After Refinement)**
He entered. Everyone stared. He said, 'Sorry, traffic was bad and it rained.' Director Chen said, 'You are late. Not good.' Zhang said, 'Okay, I fix.' Meeting continued. Problem gone.	The room goes quiet. Zhang says, '对不起陈主任，今天路上堵车，还突然下雨……' 陈主任 replies sternly, '迟到就是迟到，不要找理由。' Zhang pauses, then bows slightly: '您说得对，是我时间没安排好。我会会后整理纪要发给大家。' 陈主任 nods. Tension eases. Zhang follows through. (The room falls silent. Zhang says, "I'm sorry, Director Chen. There was heavy traffic today, and then it suddenly started raining..." Director Chen replies sternly, "Late is late—don't make excuses." Zhang pauses, then bows slightly: "You're right. I didn't manage my time properly. After the meeting, I'll prepare the minutes and share them with everyone." Director Chen nods. The tension in the room eases, and Zhang follows through on his commitment.)
Dialogue (Before Refinement)	**Dialogue (After Refinement)**
张强：抱歉，我来晚了。因为道路交通情况不好，而且天气下雨，所以我没能按时到达。 **陈主任：**（皱眉） 你迟到了，这是不对的行为。我们所有人都在等你开始会议每个人的时间都很重要。 **张强：**（低头） 是的，您说的正确。事实上，我应该更早离开家，并做出充分的出行准备。我没有合理安排时间，这是我的错误。 **陈主任：** 能够意识到错误是好的。今后请务必提前做好规划。请坐下，我们继续进行会议内容。 **张强：** 感谢您的理解。我下次会更加注意，确保不会再次发生迟到情况。 **陈主任：** 好的。希望你牢记这一点。我们现在继续讨论会议中断之前的议题。 **Zhang Qiang**: I'm sorry, I'm late. Because of heavy traffic and the rain, I couldn't arrive on time. **Director Chen**: (frowning) You were late, and that's not acceptable. Everyone was waiting for you to start the meeting. Everyone's time is important. **Zhang Qiang**: (lowering his head) Yes, you're right. In fact, I should have left home earlier and made proper travel preparations. I didn't manage my time reasonably, and that's my mistake. **Director Chen**: It's good that you can recognize your mistake. From now on, please make sure to plan ahead. Have a seat, and let's continue with the meeting. **Zhang Qiang**: Thank you for your understanding. Next time, I'll be more careful and make sure I won't be late again. **Director Chen**: Good. I hope you'll keep that in mind. Now let's continue with the agenda from before the interruption.	**张强：**（匆匆进门，微喘）陈主任，抱歉，今天路上特别堵，还突然下雨，实在没赶上时间…… **陈主任：**（皱眉，语气严肃）迟到就是迟到，别总是拿客观原因说事儿。工作要有时间观念。 （会议室一阵沉默，几位同事低头翻着资料） **张强：**（稍作停顿，鞠躬）您说得对，陈主任，是我安排不当，准备不充分，责任在我。 **陈主任：**（语气缓和）能意识到问题就好。工作就得有担当。 **张强：**我会认真反思，会议结束后我整理纪要发给大家，尽快跟上进度。 **陈主任：**行，记住别再犯同样的错误。 **张强：**明白，以后一定提前做好准备。谢谢主任的提醒。 **陈主任：**（点头）好，那我们继续会议。 **Zhang Qiang**: (hurrying in, slightly out of breath) Director Chen, I'm sorry. The traffic was really heavy today, and then it suddenly started raining, so I couldn't make it on time... **Director Chen**: (frowning, speaking sternly) Late is late. Don't always use external reasons as excuses. You need to have a proper sense of time at work. *(The meeting room falls silent, with several colleagues looking down at their documents.)* **Zhang Qiang**: (pausing briefly, bowing) You're right, Director Chen. It was poor planning and insufficient preparation on my part. The responsibility is mine. **Director Chen**: (tone softening) It's good that you can recognize the problem. Work requires taking responsibility. **Zhang Qiang**: I'll reflect on this carefully. After the meeting, I'll prepare the minutes and send them to everyone so I can quickly catch up. **Director Chen**: Alright, but remember not to make the same mistake again. **Zhang Qiang**: Understood. I'll be sure to prepare in advance next time. Thank you for the reminder, Director. **Director Chen**: (nodding) Good, then let's continue the meeting.

Figure 10: Violation-to-resolution(V2R) norm example (ZH)

Turn-Level Annotated Dialogue – V2R (EN)	Turn-Level Annotated Dialogue – V2R (KR)	Turn-Level Annotated Dialogue – V2R (ZH)
Alex: (cutting in, eyes still on notebook) Yeah, but I just think social factors are actually more relevant— **(Violation | DIS | Alex interrupts without acknowledging the speaker or waiting for their turn, showing dismissiveness and a lack of social awareness.)** **Dr. Patel:** (pauses, lifting an eyebrow, clearly caught off guard) **(Not relevant | N/A | This is a non-verbal reaction and does not constitute a verbal utterance within the labeling scheme.)** **Alex:** (keeps writing, unaware of the silence) **(Not relevant | N/A | Alex does not produce a verbal utterance in this moment; this is narrative description.)** **Dr. Patel:** (calm but firm) Alex, hold on—I hadn't finished my point. **(Adhered | CRT | Dr. Patel gives a calm correction, pointing out the interruption respectfully, aligning with appropriate academic authority.)** **Alex:** (looks up, noticing the shift in tone) Oh—sorry, Dr. Patel. I didn't mean to jump in like that. **(Adhered | APO | Alex issues an immediate and direct apology after realizing the inappropriate behavior.)** **Dr. Patel:** (slightly frowning) I know you're eager to contribute, and that's great. But let's make sure we're giving everyone space to finish their thoughts. **(Adhered | SUG | Dr. Patel acknowledges Alex's intent and offers a constructive suggestion on classroom discussion norms.)** **Alex:** (face flushing, now fully engaged) You're right. I apologize—I got ahead of myself. Please, go ahead. **(Adhered | APO | Alex reiterates the apology with emotional awareness and explicitly yields the floor.)** **Dr. Patel:** (nodding, tone softens) Thanks, Alex. I appreciate that. We're all here to learn from each other, and that works best when we listen, too. **(Adhered | THX | Dr. Patel acknowledges the apology and affirms a collaborative learning environment.)** **Alex:** (genuinely) Absolutely. I'll be more mindful going forward. Really sorry again. **(Adhered | APO | Alex again apologizes and commits to improved behavior, showing self-awareness and respect.)** **Dr. Patel:** (smiling) No worries. Let's pick up where we left off… **(Adhered | EMP | Dr. Patel offers emotional support and diffuses tension, signaling forgiveness and readiness to move on.)** **Alex:** (sits up straighter, fully focused) **(Not relevant | N/A | This is non-verbal behavior and not an utterance to be labeled.)**	**사장님:** 민수 씨, 이번 매출 보고서에서 지난달 수치가 잘못 기재된 것 같네요. 확인해 보셨어요? **(Adhered | QUE | 정중한 어투로 문제 상황을 지적하며 확인 요청함)** **President**: Minsu, it looks like last month's figures were entered incorrectly in this sales report. Have you checked it? (Adhered | QUE | Politely points out the issue and requests confirmation.) **민수:** (급하게 모니터를 보며) 아, 네… 아마 회계팀 자료가 중간에 바뀌어서 그럴 거예요. 저도 헷갈려서… **(Violation | JUS | 문제를 인정하기보다 다른 팀 탓으로 돌리며 책임을 회피하려는 정당화 시도)** **Minsu**: (quickly looking at the monitor) Oh, yes… It's probably because the accounting team changed the data midway. I got confused too… (Violation | JUS | Attempts to justify by shifting blame to another team rather than taking responsibility.) **사장님:** (표정이 딱딱해지며) 다른 팀 때문이라는 건가요? **(Adhered | QUE | 책임 회피로 보이는 민수의 발언에 대해 재확인하며 의도 파악을 시도)** **President**: (expression hardening) Are you saying it was the other team's fault? (Adhered | QUE | Seeks clarification of Minsu's seemingly evasive statement.) **민수:** (잠깐 정적, 민수는 사장님의 표정을 보고 잠시 생각한다) **(Not relevant | N/A | 내적 반성과 상황 관찰을 묘사한 내레이션으로, 발화 없음)** **Minsu**: (a brief silence, as he notices the president's expression and reflects for a moment) (Not relevant | N/A | Narration of inner reflection and observation; no actual utterance.) **민수:** 아… 죄송합니다, 사장님. 변명을 의도하려한 건 아닌데, 잘못을 제대로 인정하지 않고 말씀드린 점, 정말 사과드립니다. 제 실수입니다. **(Adhered | APO | 본인의 실수를 인정하고 진심 어린 사과를 명확히 표현함)** **Minsu**: Ah… I'm sorry, President. I didn't mean to make excuses, but I realize I didn't fully acknowledge the mistake when I spoke. I sincerely apologize. This is my error. (Adhered | APO | Clearly admits his mistake and offers a genuine apology.) **사장님:** (조금 부드러워진 목소리로) 그래요. 누구나 실수할 수는 있어요. 다음엔 좀 더 꼼꼼히 확인해 주세요. 지금 수정은 가능합니까? **(Adhered | EMP | 실수 가능성을 인정하며 공감 표현 + 실질적 해결을 위한 제안 포함)** **President**: (voice softening slightly) Alright. Anyone can make mistakes. Just check more carefully next time. Can it be corrected now? (Adhered | EMP | Expresses empathy by acknowledging the possibility of mistakes, while also proposing a practical solution.) **민수:** 네, 바로 수정해서 보내드리겠습니다. 다시 한 번 죄송합니다. **(Adhered | APO | 책임을 수용하고 빠른 조치를 약속하며 재차 사과함)** **Minsu**: Yes, I'll correct it right away and send it to you. I'm truly sorry again. (Adhered | APO | Accepts responsibility, promises prompt action, and apologizes once more.) **사장님:** 앞으로는 문제가 생기면 바로 인정하고, 해결책도 같이 고민하도록 해요. 그게 더 신뢰를 줄 수 있어요. **(Adhered | SUG | 문제 해결 방식을 제시하며 조언과 신뢰 형성에 대한 가치를 전달함)** **President**: From now on, if a problem arises, admit it immediately and work on a solution together. That way, you'll build more trust. (Adhered | SUG | Provides guidance on how to handle mistakes and emphasizes the value of trust.) **민수:** 네, 명심하겠습니다. 감사합니다. **(Adhered | ACK + THX | 조언을 수용하고 감사 인사를 전함)** **Minsu**: Yes, I'll keep that in mind. Thank you. (Adhered | ACK + THX | Accepts the advice and expresses gratitude.) (잠시 후 회의실 분위기가 조금 누그러지고, 민수는 서둘러 수치를 다시 확인한다. 사장님도 다음 안건으로 넘어간다. 민수는 마음을 다잡으며 업무에 집중한다.) **(Not relevant | N/A | 대화 외 내면 묘사 및 서술)** (A moment later, the atmosphere in the meeting room eases. Minsu quickly reviews the figures, while the president moves on to the next agenda item. Minsu steadies himself and refocuses on his work.) (Not relevant | N/A | Narrative description of the scene rather than spoken dialogue.)	**张强:** （匆匆进门，微喘）陈主任, 抱歉, 今天路上特别堵，还突然下雨，实在没赶上时间…… **(Violation | JUS | 张强 is providing external reasons (traffic, rain) for his tardiness, which reflects a justification and not a direct apology.)** **Zhang Qiang**: (hurrying in, slightly out of breath) Director Chen, I'm sorry. The traffic was really heavy today, and it suddenly started raining, so I just couldn't make it on time… **陈主任:** （皱眉，语气严肃）迟到就是迟到，别总是拿客观原因说事儿。工作要有时间观念。 **(Adhered | CRT | 陈主任 criticizes the use of excuses and emphasizes the importance of time management, reflecting proper professional norm enforcement.)** **Director Chen**: (frowning, speaking sternly) Late is late. Don't always blame external reasons. You need to have a sense of time in your work. （会议室一阵沉默，几位同事低头翻着资料） **(Not relevant | N/A | This is a narrative action description, not a verbal utterance.)** *(The meeting room falls silent, and several colleagues lower their heads, flipping through their materials.)* **张强:**（稍作停顿，鞠躬）您说得对，陈主任，是我安排不当，准备不充分，责任在我。 **(Adhered | APO | 张强 accepts responsibility without further excuses and acknowledges his own fault—this is a culturally appropriate apology.)** **Zhang Qiang**: (pausing briefly, bowing) You're right, Director Chen. It was poor planning and insufficient preparation on my part. The responsibility is mine. **陈主任:**（语气缓和）能意识到问题就好。工作就得有担当。 **(Adhered | AGR | 陈主任 softens his tone and expresses agreement that recognizing one's mistake is good; he reinforces responsibility.)** **Director Chen**: (tone softening) It's good that you can recognize the problem. Work requires a sense of responsibility. **张强:** 我会认真反思，会议结束后我整理纪要发给大家，尽快跟上进度。 **(Adhered | SUG | 张强 suggests a concrete follow-up action (writing and sharing meeting minutes) to take responsibility and make amends.)** **Zhang Qiang**: I will seriously reflect on this. After the meeting, I'll prepare and share the minutes with everyone so I can catch up quickly. **陈主任:** 行，记住别再犯同样的错误。 **(Adhered | SUG | 陈主任 provides advice and a warning not to repeat the same mistake—norm-consistent correction from a superior.)** **Director Chen**: Alright, just remember not to make the same mistake again. **张强:** 明白，以后一定提前做好准备。谢谢主任的提醒。 **(Adhered | THX | 张强 acknowledges the guidance and expresses gratitude, showing politeness and willingness to improve.)** **Zhang Qiang**: Understood. I'll make sure to prepare in advance from now on. Thank you for the reminder, Director. **陈主任:**（点头）好，那我们继续会议。 **(Adhered | N/A | This utterance resumes the meeting and closes the interaction; no specific dialog act beyond transition.)** **Director Chen**: (nodding) Good, then let's continue the meeting.

Figure 11: Turn-level annotated dialogue example – V2R

Category	Subnorm
Apology	Bowing to elders when apologizing is not considered necessary or meaningful. In American culture, apologies are not typically accompanied by physical gestures like bowing, even when addressing older or higher-status individuals. What matters more is the sincerity conveyed through eye contact, tone, and word choice. For example, one might say, "I'm really sorry, sir. I take full responsibility," while maintaining a calm and respectful demeanor.

Scenario(After Refinement)

Between classes in a busy high school hallway, Amy, a sophomore, accidentally steps on the foot of the school principal, Mr. Carter, while hurrying past with her friends.

Situation(After Refinement)

The hallway was crowded with students rushing to their next class. Amy, chatting and laughing with her friends, wasn't watching where she was going. Suddenly, she stepped hard on something. Looking down, she realized it was someone's foot. She looked up—and froze. It was Principal Carter. Her face flushed, but instead of apologizing, she let out a quick laugh and muttered to her friends, "Oops, guess I should watch where I'm going." Without making eye contact, she tugged her backpack strap and kept walking, still giggling. Principal Carter stood still, raising an eyebrow as he watched her disappear around the corner—without so much as a word.

Dialogue (After Refinement)

Amy: (laughing, bumping into someone) Oh man, did you see that?
Principal Carter: (stopping, rubbing his foot) Excuse me—
Amy: (glancing up, recognizing him, then turning back to her friends) Ugh, that was close, I almost ate it.
Principal Carter: (expectant, watching Amy) Amy, you just—
Amy: (interrupting, not making eye contact) Yeah, sorry guys, I'm like half asleep today. C'mon, let's go, we're gonna be late.
Principal Carter: (eyes narrowing, voice firm) Amy—
Amy: (already walking away with friends, over her shoulder) See ya, Mr. Carter!
Principal Carter: (standing still, sighs quietly, watching her walk off)

Figure 12: Violation norm example (EN)

Relevance Prompt

Section	Content
Evaluation Instruction	You are a professional dataset auditor for social-norm dialogues. You are given a culture category. Your task is to evaluate only the *Relevance* of the dialogue to the provided Scenario and Situation. Ignore grammar, fluency, or logical consistency. Focus on whether the dialogue reflects the key intentions, emotions, and context described in the Scenario and Situation.
Parameter	`culture`, `scenario`, `situation`, `dialogue`
Evaluation Question	Does the dialogue appropriately address and reflect the actions, intentions, and emotional context presented in the scenario and situation? – Are the key elements of the situation represented in the conversation (e.g., apology, embarrassment, disagreement)? – Do the characters react in a way that makes sense for the described context? – Are any critical actions or emotional responses missing from the dialogue?
Scoring Criteria	1 = Dialogue is mostly irrelevant to the situation 3 = Partially relevant, with some elements missing or misaligned 5 = Dialogue is highly relevant and faithfully represents the described situation

Table 25: Evaluation prompt structure for *Relevance*.

Category	Subnorm
Apology	윗사람에게 사과할 때는 '변명 없이 바로 사과'하는 것이 중요하게 여겨진다. (When apologizing to a superior, it is considered important to apologize directly without making excuses.)
Scenario(Before Refinement)	**Scenario(After Refinement)**
민수는 회의 시간 중 발표 도중 잘못된 숫자를 언급했으나, 사장님이 질문했을 때 즉시 사과하지 않고 복잡한 이유들과 설명을 계속하며 자신의 잘못이 아닌 듯한 인상을 주려고 한다. (During a meeting, Minsu mentioned an incorrect number in his presentation. However, when the president questioned him, instead of apologizing immediately, he kept giving complicated reasons and explanations, creating the impression that the mistake was not his fault.)	민수는 회의 중에 잘못된 정보를 발표했지만, 사장님이 그 내용을 지적했을 때 바로 사과하지 않고 여러 가지 변명을 하며 자신의 실수를 인정하지 않는다. (Minsu presented incorrect information during the meeting, but when the president pointed it out, he did not apologize immediately and instead made various excuses, refusing to acknowledge his mistake.)
Situation(Before Refinement)	**Situation(After Refinement)**
민수는 회의실에서 다소 빠른 말투로 매출 데이터를 발표하고 앉았다. 사장님이 그제서야 조용히 "민수 씨, 이거 지난 분기 숫자랑 다르네요, 무슨 사정이죠?"라고 물었다. 민수는 고개를 왼쪽으로 살짝 기울이고 눈을 빠르게 깜빡이며, "아, 네... 그게 말이죠... 음, 사실은 제가 그 숫자를 받은 게 어제 저녁이었고요, 다른 팀이 뭘 보내줬는데 정확한 수치가 없었어요. 그리고 제가 밤새 정리하다가 아침에 다시 보고... 네..."라고 횡설수설했다. 그는 억지로 웃으며 "그래서 조금 어긋났을 수도 있는데, 그래도 큰 차이는 아니니까 괜찮을 것 같아요"라고 덧붙였다. 사장님은 무표정하게 고개를 끄덕이거나 말거나 하며 침묵했고, 회의실은 멍한 공기 속에 잠겼다. 민수는 그 뒤에도 몇 마디 더 중얼거리듯 이야기했지만, 아무도 반응하지 않았다. (In the meeting room, Minsu delivered the sales data in a rather quick tone and then sat down. The president quietly asked, “Minsu, these numbers are different from last quarter’s. What’s the reason?” Minsu tilted his head slightly to the left, blinked rapidly, and stammered, “Ah, yes... well, you see... um, actually I only received those numbers last night, and another team sent something without precise figures. Then I stayed up all night trying to sort it out, and looked at it again this morning... yes...” Forcing a smile, he added, “So there may be a slight discrepancy, but it’s not a big difference, so I think it should be fine.” The president nodded blankly, expressionless, and remained silent, leaving the room in a dazed atmosphere. Minsu mumbled a few more words afterward, but no one responded.)	회의실에서 민수는 팀원들 앞에서 회사의 매출 수치를 발표했다. 발표가 끝난 후 사장님이 "민수 씨, 이 자료의 숫자가 지난 분기와 맞지 않는데, 무슨 일인가요?"라고 물었다. 민수는 순간 당황한 표정을 지었지만 곧 침착한 척하며 말했다. "아, 그게... 사실 어제까지 자료가 계속 업데이트되고 있어서요. 그리고 다른 부서에서도 자료를 늦게 줘서 정확히 맞추기가 어려웠습니다." 그는 살짝 미소를 지으며 "자료가 완전히 정리되지 않은 상황이었어요"라고 덧붙였다. 사장님의 표정이 굳어졌지만, 민수는 자세한 해명을 이어갔고, 명확한 사과의 말은 하지 않았다. 회의실에는 어색한 정적이 흘렀다. (In the meeting room, Minsu presented the company’s sales figures in front of his team. After the presentation, the president asked, “Minsu, these numbers don’t match last quarter’s. What happened?” Minsu looked momentarily flustered but quickly pretended to stay composed and said, “Ah, well... actually, the data kept being updated until yesterday. And another department sent their figures late, so it was difficult to match everything exactly.” Smiling faintly, he added, “The data just wasn’t fully organized yet.” The president’s expression hardened, but Minsu continued with detailed explanations without offering a clear apology. An awkward silence filled the meeting room.)
Dialogue (Before Refinement)	**Dialogue (After Refinement)**
사장님: 민수 씨, 여기 매출 데이터가 지난 분기랑 맞지 않는 것 같은데, 무슨 일인가요? **민수:** 아, 네... 그게, 사실 어제 저녁에 자료를 받아서요. 다른 팀에서도 정확하게 자료를 안 보내주셔서, 좀 헷갈렸어요. 밤새 정리하느라 아침에 다시 한 번 보긴 했는데, 음... 네, 그래서 살짝 다를 수도 있거든요. **사장님:** (잠시 정적) 음... 그래요. **민수:** 하하, 뭐, 그렇게 큰 차이는 아니라고 생각합니다. 아무래도 팀에서 주신 자료가 좀 늦어서... 네, 아무튼 다음엔 아마도 더 잘 맞출 수 있을 겁니다. **사장님:** (무표정) 알겠습니다. **민수:** 네... (작게 중얼거리며) 음, 뭐... 팀마다 방식이 달라서요. **President**: Minsu, these sales figures don’t seem to match last quarter’s. What happened? **Minsu**: Ah, yes... well, I actually received the data only last night. Another team didn’t send their numbers accurately, so I got a bit confused. I stayed up all night organizing it and looked over it again this morning, um... yes, so it might be slightly different. **President**: (after a brief pause) Hmm... I see. **Minsu**: Haha, well, I don’t think it’s a big difference. It’s just that the team’s data came in a bit late... Anyway, next time I’ll probably be able to match it better. **President**: (expressionless) Alright. **Minsu**: Yes... (muttering softly) Well... every team has a different way of doing things.	**사장님:** 민수 씨, 이 매출 수치가 지난 분기랑 안 맞는 것 같은데, 이유가 뭐죠? **민수:** 아, 네. 그게... 어제까지 자료가 계속 바뀌고 있어가지고요. 다른 부서에서도 자료 넘겨주는데 좀 늦었습니다. **사장님:** 그래도 발표 자료인데, 이런 부분은 정확해야 하지 않나요? **민수:** 네, 근데 자료가 완전히 정리가 안 됐던 상황이었어요. 사실 시간이 너무 촉박해서 좀 힘들었습니다. **사장님:** (잠시 침묵) 민수 씨, 이런 실수는 앞으로 신경 좀 써줬으면 해요. **민수:** 네, 알겠습니다. 다음엔 더 신경 쓸게요. 어제도 부서에서 자료 오는 게 늦어서 좀 곤란했거든요. (회의실 분위기가 무거워지고, 모두 민수의 말을 조용히 듣는다) **President**: Minsu, these sales figures don’t seem to match last quarter’s. What’s the reason? **Minsu**: Ah, yes. Well... the data kept changing until yesterday. And another department was late in sending their part as well. **President**: Still, this is presentation material. Shouldn’t details like this be accurate? **Minsu**: Yes, but the data wasn’t fully organized. Honestly, the deadline was very tight, so it was difficult. **President**: (after a brief silence) Minsu, I’d like you to be more careful about this kind of mistake in the future. **Minsu**: Yes, understood. I’ll be more careful next time. It’s just that yesterday the department’s data came in late, which made it difficult. *(The atmosphere in the meeting room grows heavy, and everyone listens quietly to Minsu’s words.)*

Figure 13: Violation norm example (KR)

Category	Subnorm
Apology	When apologizing to someone in a higher position, offering excuses beforehand is typically seen as inappropriate in Chinese culture. A direct assumption of responsibility is more acceptable and considered polite. For instance, saying “这次是我的责任，没有任何借口” immediately shows sincerity and self-awareness without appearing defensive or disrespectful.
Scenario(Before Refinement)	**Scenario(After Refinement)**
Chen Jie came ten minutes late. Manager Liu stopped talking. She didn’t say sorry.	During a team meeting, 陈洁 comes in 10 minutes late. 刘经理 pauses. She explains instead of apologizing.
Situation(Before Refinement)	**Situation(After Refinement)**
She looked tired, holding her bag. 'Manager Liu, the road was very blocked and subway stopped.' Nobody spoke. It was quiet. She sat, moved her stuff. The meeting continued, but felt weird.	Tuesday morning. 陈洁 enters, out of breath. '刘经理，今天路上太堵了，地铁也停了，不是故意迟到的。' Silence. 刘经理 says nothing. Colleagues exchange looks. 陈洁 avoids eye contact, fidgets. The meeting resumes, but tension remains. (Tuesday morning. Chen Jie enters, out of breath. “Manager Liu, the traffic was terrible today, and the subway broke down—I really didn’t mean to be late.” Silence. Manager Liu says nothing. Colleagues exchange glances. Chen Jie avoids eye contact, fidgeting. The meeting resumes, but the tension lingers in the room.)
Dialogue (Before Refinement)	**Dialogue (After Refinement)**
张强：抱歉，我来晚了。因为道路交通情况不好，而且天气下雨，所以我没能按时到达。 **陈主任：**（皱眉） 你迟到了，这是不对的行为。我们所有人都在等你开始会议每个人的时间都很重要。 **张强：**（低头） 是的，您说的正确。事实上，我应该更早离开家，并做出充分的出行准备。我没有合理安排时间，这是我的错误。 **陈主任：** 能够意识到错误是好的。今后请务必提前做好规划。请坐下，我们继续进行会议内容。 **张强：** 感谢您的理解。我下次会更加注意，确保不会再次发生迟到情况。 **陈主任：** 好的。希望你牢记这一点。我们现在继续讨论会议中断之前的议题。 **Zhang Qiang**: I’m sorry I’m late. Because of the bad traffic conditions and the rain, I couldn’t arrive on time. **Director Chen**: (frowning) You were late, and that’s not acceptable. Everyone was waiting for you to start the meeting. Everyone’s time is important. **Zhang Qiang**: (lowering his head) Yes, you’re right. In fact, I should have left home earlier and made proper travel preparations. I didn’t manage my time well—this was my mistake. **Director Chen**: It’s good that you can recognize your mistake. From now on, please plan ahead. Have a seat, and let’s continue with the meeting. **Zhang Qiang**: Thank you for your understanding. Next time, I’ll be more careful and make sure I won’t be late again. **Director Chen**: Good. I hope you keep that in mind. Now, let’s continue with the agenda from before the interruption.	**张强：**（匆匆进门，微喘）陈主任，抱歉，今天路上特别堵，还突然下雨，实在没赶上时间…… **陈主任：**（皱眉，语气严肃）迟到就是迟到，别总是拿客观原因说事儿。工作要有时间观念。 （会议室一阵沉默，几位同事低头翻着资料） **张强：**（稍作停顿，鞠躬）您说得对，陈主任，是我安排不当，准备不充分，责任在我。 **陈主任：**（语气缓和）能意识到问题就好。工作就得有担当。 **张强：** 我会认真反思，会议结束后我整理纪要发给大家，尽快跟上进度。 **陈主任：** 行，记住别再犯同样的错误。 **张强：** 明白，以后一定提前做好准备。谢谢主任的提醒。 **陈主任：**（点头）好，那我们继续会议。 **Zhang Qiang**: (hurrying in, slightly out of breath) Director Chen, I’m sorry. The traffic was terrible today, and then it suddenly started raining—I really couldn’t make it on time... **Director Chen**: (frowning, stern tone) Late is late. Don’t always use external reasons as excuses. You need to have a proper sense of time at work. *(The meeting room falls silent, with several colleagues looking down at their documents.)* **Zhang Qiang**: (pausing briefly, bowing) You’re right, Director Chen. It was poor planning and insufficient preparation on my part. The responsibility is mine. **Director Chen**: (tone softening) It’s good that you can recognize the problem. Work requires taking responsibility. **Zhang Qiang**: I’ll reflect on this carefully. After the meeting, I’ll prepare the minutes and share them with everyone so I can catch up quickly. **Director Chen**: Alright, but remember not to make the same mistake again. **Zhang Qiang**: Understood. Next time, I’ll be sure to prepare in advance. Thank you for the reminder, Director. **Director Chen**: (nodding) Good, then let’s continue the meeting.

Figure 14: Violation norm example (ZH)

Turn-Level Annotated Dialogue – Violation (EN)	Turn-Level Annotated Dialogue – Violation (KR)	Turn-Level Annotated Dialogue – Violation (ZH)
Amy: (laughing, bumping into someone) Oh man, did you see that? (Violation | N/A | Amy casually reacts to bumping into someone without checking who it is or acknowledging responsibility.) **Principal Carter**: (stopping, rubbing his foot) Excuse me— (Adhered | ACK | Principal Carter expresses having been affected and implicitly expects acknowledgment.) **Amy**: (glancing up, recognizing him, then turning back to her friends) Ugh, that was close, I almost ate it. (Violation | DIS | Amy notices that it's the principal but ignores the social expectation to apologize, instead making a dismissive comment.) **Principal Carter**: (expectant, watching Amy) Amy, you just— (Adhered | QUE | Principal Carter begins to address the incident, likely to elicit accountability or a response.) **Amy**: (interrupting, not making eye contact) Yeah, sorry guys, I'm like half asleep today. C'mon, let's go, we're gonna be late. (Violation | JUS | Amy gives an excuse to her friends without properly addressing Principal Carter or apologizing to him.) **Principal Carter**: (eyes narrowing, voice firm) Amy— (Adhered | CRT | The principal attempts to address her inappropriate behavior again, showing disapproval.) **Amy**: (already walking away with friends, over her shoulder) See ya, Mr. Carter! (Violation | DIS | Amy dismissively ends the interaction without resolving the situation, signaling disregard.) **Principal Carter**: (standing still, sighs quietly, watching her walk off) (Not relevant | N/A | A descriptive action that shows disappointment but is not a verbal utterance.)	**사장님**: 민수 씨, 이 매출 수치가 지난 분기랑 안 맞는 것 같은데, 이유가 뭐죠? **(Adhered | QUE | 사장님이 문제 상황에 대해 설명을 요청하며, 정중하게 질문함)** **President**: Minsu, these sales figures don't seem to match last quarter's. Can you explain why? (Adhered | QUE | The president politely requests clarification regarding the issue.) **민수**: 아, 네. 그게… 어제까지 자료가 계속 바뀌고 있어가지고요. 다른 부서에서도 자료 넘겨주는데 좀 늦었습니다. **(Violation | JUS | 민수가 실수를 직접 인정하기보다 외부 요인을 언급하며 정당화를 시도함)** **Minsu**: Ah, yes. Well… the data kept changing until yesterday. Another department was supposed to send it over, but they were late. (Violation | JUS | Instead of directly admitting his mistake, Minsu cites external factors as justification.) **사장님**: 그래도 발표 자료인데, 이런 부분은 정확해야 하지 않나요? **(Adhered | CRT | 사장님이 업무 책임감을 강조하며 정확성 부족을 지적함)** **President**: Still, this is presentation material. Shouldn't this kind of detail be accurate? (Adhered | CRT | The president stresses accountability and points out the lack of accuracy.) **민수**: 네, 근데 자료가 완전히 정리가 안 됐던 상황이었어요. 사실 시간이 너무 촉박해서 좀 힘들었습니다. **(Violation | JUS | 민수가 또다시 상황 탓을 하며 자신의 책임을 명확히 인정하지 않고 정당화함)** **Minsu**: Yes, but the data wasn't fully finalized yet. Honestly, the deadline was too tight, so it was difficult. (Violation | JUS | Minsu again blames circumstances, failing to clearly acknowledge his own responsibility.) **사장님**: (잠시 침묵) 민수 씨, 이런 실수는 앞으로 신경 좀 써줬으면 해요. **(Adhered | SUG | 사장님이 직접적인 비난 없이 개선을 촉구하며 조언을 제시함)** **President**: (pausing) Minsu, I'd like you to be more careful about this kind of mistake in the future. (Adhered | SUG | The president offers advice and a suggestion for improvement without direct blame.) **민수**: 네, 알겠습니다. 다음엔 더 신경 쓸게요. 어제도 부서에서 자료 오는 게 늦어서 좀 곤란했거든요. **(Violation | JUS | 처음에 인지를 표현했으나, 바로 이어서 다시 타부서 탓을 하며 책임을 흐림)** **Minsu**: Yes, I understand. I'll be more careful next time. It was just that the department sent the data late yesterday, which made things difficult. (Violation | JUS | While first acknowledging the advice, Minsu immediately shifts blame again, weakening his responsibility.) (회의실 분위기가 무거워지고, 모두 민수의 말을 조용히 듣는다) **(Not relevant | N/A | 내레이션으로 발화가 아니며, 라벨링과 무관함)** (The atmosphere in the meeting room grows heavy, and everyone listens quietly to Minsu's words.) (Not relevant | N/A | This is narration, not spoken dialogue, and thus not part of the labeling.)	**陈洁:** (急匆匆推门进来，喘着气) 刘经理，今天路上太堵了，那个地铁还突然坏了，真不是故意迟到的。 **(Violation | JUS | Chen Jie does not offer a direct apology but instead justifies her tardiness by blaming external factors like traffic and subway failure.)** **Chen Jie**: (pushing the door open hurriedly, out of breath) Manager Liu, the traffic was terrible today, and then the subway suddenly broke down. I really didn't mean to be late. **刘经理:** (抬眼看了她一眼，语气平淡) 嗯，大家都等了你一会儿，咱们现在开始吧。 **(Adhered | CRT | Manager Liu indirectly criticizes her lateness by pointing out the group had to wait, then moves forward professionally.)** **Manager Liu**: (looking up at her, tone flat) Hmm, everyone's been waiting for you for a while. Let's begin now. **陈洁:** (低头放下包，局促地坐下) 哎，每次都是早出门，结果碰上这种事也没办法。 **(Violation | JUS | Instead of acknowledging the inconvenience caused or apologizing, Chen Jie continues to justify her lateness.)** **Chen Jie**: (lowering her head, setting down her bag awkwardly, and sitting) Sigh, every time I leave early, something like this happens. There's nothing I can do. **刘经理:** (翻开手里的资料，没有回应) 现在我们说一下本周的项目进度…… **(Adhered | N/A | Manager Liu ignores her justification and returns to the meeting agenda, showing professionalism without engaging further.)** **Manager Liu**: (opening the materials in his hand without responding) Now, let's go over this week's project progress… **陈洁:** (小声嘀咕) 本来就不是我的错，真倒霉。 **(Violation | JUS | Chen Jie mutters a complaint, continuing to deny responsibility and reinforcing the lack of self-awareness.)** **Chen Jie**: (muttering quietly) It wasn't my fault anyway… just bad luck. **刘经理:** (停顿片刻，略显不悦地扫了一眼) 那我们先从市场组汇报开始。 **(Adhered | ACK | Manager Liu responds with a neutral action to move the meeting forward, subtly signaling disapproval but not escalating.)** **Manager Liu**: (pausing briefly, glancing at her with slight displeasure) Let's start with the market team's report. **陈洁:** (抿嘴，继续整理文件，避免和刘经理对视) **(Not relevant | N/A | Non-verbal behavior with no verbal utterance; no direct label applies.)** **Chen Jie**: (pressing her lips together, organizing her files, avoiding eye contact with Manager Liu) (会议室里气氛有些压抑，同事们彼此对视，但没人说话，会议照常进行，尴尬和紧张仍然萦绕在空气中) **(Not relevant | N/A | Narrative description without dialogue helps set the tone but is not a labelable utterance.)** (The atmosphere in the meeting room feels somewhat tense. Colleagues glance at each other, but no one speaks. The meeting proceeds as usual, though the air is filled with awkwardness and unease.)

Figure 15: Turn-level annotated dialogue example – Violation

Emotional Appropriateness Prompt

Section	Content
Evaluation Instruction	You are a professional dataset auditor for social-norm dialogues. You are given a culture category. Your task is to evaluate only the *Emotional Appropriateness* of the dialogue. Ignore grammar, norm correctness, or logical structure. Focus on whether the tone, expressions, and emotional language used in the dialogue match the emotional context of the situation.
Parameter	`culture`, `scenario`, `situation`, `dialogue`
Evaluation Question	Does the emotional tone, choice of words, and manner of speaking in the dialogue align appropriately with the emotional context of the situation? – Does the dialogue reflect the expected emotional state (e.g., tension, regret, embarrassment, relief) implied in the situation? – Are the expressions and tone suitable for the described emotional stakes? – Is there any emotional mismatch that makes the dialogue feel unnatural or inappropriate?
Scoring Criteria	1 = Emotionally disconnected or inappropriate 3 = Emotion is somewhat present but weak or inconsistent 5 = Emotional tone is highly appropriate and enhances the realism

Table 26: Evaluation prompt structure for *Emotional Appropriateness*.

Social Norm Appropriateness Prompt

Section	Content
Evaluation Instruction	You are a professional dataset auditor for social-norm dialogues. You are given a culture category. Your task is to evaluate only the *Social Norm Appropriateness* of the dialogue. Assess how well the conversation reflects the given social norm, and categorize the degree of adherence.
Parameter	`culture`, `norm`, `dialogue`
Evaluation Question	Based on the given social norm, how well does the dialogue align with it? – Does the dialogue completely follow the norm? – Does it violate the norm? – Does it violate the norm but later attempt to resolve it? – Is the behavior partially aligned with the norm?
Classification Criteria	1 = Fully Violated 2 = Partially Violated 3 = Violation then Resolved 4 = Partially Adherence 5 = Fully Adherence

Table 27: Evaluation prompt structure for *Social Norm Appropriateness*.

Scenario Coherence Prompt

Section	Content
Evaluation Instruction	You are a professional dataset auditor for social-norm dialogues. Your task is to evaluate only the *Scenario Coherence* of the dialogue. Ignore grammar or emotional tone. Focus on whether the dialogue logically and smoothly follows the sequence of events described in the scenario.
Parameter	`culture`, `scenario`, `situation`, `dialogue`
Evaluation Question	Does the dialogue unfold in a way that aligns with the narrative structure and event flow of the scenario? – Is there a smooth and coherent transition from the described situation into the dialogue? – Are there any gaps, jumps, or inconsistencies between what the scenario sets up and what happens in the conversation? – Does the dialogue logically follow the order of events and emotional pacing described in the scenario (or situation)?
Scoring Criteria	1 = Dialogue and scenario (or situation) are disconnected or contradictory 3 = Some transitions or event links are missing or unclear 5 = Dialogue flows logically and coherently from the scenario (or situation)

Table 28: Evaluation prompt structure for *Scenario Coherence*.

Task	Prompt Template
Continuation Generation	You are continuing the following human-human dialogue. Continue the conversation with 5 additional turns naturally and coherently. Dialogue so far: {context} Continue the dialogue:
A/B Evaluation	You are an evaluator. Below is a dialogue context followed by two different continuations (A and B). Choose which continuation is more appropriate, natural, coherent, and socially norm-aligned. Respond only with "A" or "B". Context: {context} Response A: {output_a} Response B: {output_b} Which is better? Respond with A or B only.

Table 29: Prompt templates used for generalization quality evaluation.

Example 1. Lexical Redundancy

늦은 밤, 거실에서 혼자 책을 읽던 아버지 앞에 태훈이 조심스럽게 앉았다. 최근 면접 탈락 소식을 들고도 태훈은 아버지께 바로 말하지 못했다.
태훈은 **고개를 숙이며 말했다.**
"아버지…, 저 사실 면접 떨어졌어요. 말씀드리기 힘들었어요."
아버지는 책을 내려놓고 천천히 고개를 들었다.
태훈은 **고개를 숙이며 말했다.** "기대를 저버린 것 같아서… 너무 죄송해요. 실망하셨죠?"
아버지는 잠시 침묵하다가, "태훈아, 아버지가 바라는 건 결과보다 너의 노력이라는 거… 알지?"
태훈은 눈시울이 붉어졌다. 태훈은 **고개를 숙이며 말했다.** "앞으로 더 나아지는 모습 보여드릴게요. 다시 시작할게요."
그 옆에 어머니가 조용히 미소를 지었다.

Late at night, Taehun cautiously sat in front of his father, who was reading alone in the living room. Even after hearing the recent news of his test failure, Taehun could not immediately speak to his father.
Taehun said in a trembling voice.
"Father… I actually failed the interview. It was hard for me to tell you."
His father put down the book and quietly listened.
Taehun said in a trembling voice.
"I feel like I let you down… I'm so sorry. You must be disappointed, right?"
His father paused for a moment, then said softly,
"Taehun, what your father wants is not the result, but your effort… You know that, don't you?"
Tears welled up in Taehun's eyes. He said in a trembling voice,
"I'll show you a better version of myself from now on. I'll start again."
Next to them, his mother smiled quietly.

During a lively seminar, Alex, eager to share a point, jumps in and cuts off Dr. Patel mid-sentence, saying, "Yeah, but I think it's more about social factors," without looking up from their notes. The room falls quiet for a moment. Dr. Patel raises her eyebrows slightly, caught off guard. Realizing the interruption and how dismissive it sounded, Alex pauses, then looks up to meet her eyes. "I'm sorry, Dr. Patel—I didn't mean to interrupt. Please go ahead," they say sincerely. Dr. Patel's expression softens, and she gives a small nod before continuing. The tension in the room eases, and Alex listens more attentively, relieved to have handled the moment with maturity.

Example 2. Tone Mismatches

저녁 약속 시간보다 한참 늦은 시각, 민수가 하동진's 카페에 도착했다. 이미 친구들은 테이블에 앉아 음료를 마시며 그를 기다리고 있었다.
"얘들아, 정말 **미안해요.** 시간을 오후 7시로 착각했어. 너희 다 기다리게 했지?"
민수는 숨을 고르며 고개를 숙였고, 지수는 웃으며 "뭐, 별거 아니니까. 괜찮아. 이제 왔잖아."
다른 친구들도 "우리 아직 메뉴도 안 시켰어. 네가 기다리느라 고생했네." 하고 말하며 분위기를 풀었다.
민수는 숨을 모으며 "다음부터 진짜 캘린더 잘 확인할게. 내가 오늘 디저트 쏠게."
그 말에 친구들은 웃음을 터뜨렸고, 민수는 안도의 미소를 지었다.

It was much later than the scheduled dinner time when Minsu arrived at Hadongjin's café. His friends were already sitting at a table, drinking and waiting for him.
"Guys, I'm really sorry. I thought the time was 7 p.m. I must have kept you all waiting."
Catching his breath, Minsu bowed his head. Jisoo smiled and said, "Well, it's no big deal. It's fine. You're here now."
Another friend added, "We haven't even ordered yet. You must have had a hard time rushing over." Lightening the mood.
Taking a deep breath, Minsu said, "I'll make sure to double-check my calendar next time. Dessert's on me today."
At that, his friends burst into laughter, and Minsu smiled in relief.

We observe a stylistic inconsistency in the generated dialogue where the speaker mixes honorific and casual forms within the same utterance or across adjacent turns. In particular, phrases like "미안해요" (formal) and "얘들아, 기다리게 했지?" (informal) co-occur, despite the setting involving close friends. This mismatch in speech levels is especially problematic in Korean, where speech style directly encodes social relationships and formality. Such inconsistencies often arise from token-level decoding that lacks discourse-level awareness of speaker roles and relational context. Addressing this issue is essential for generating culturally coherent and socially appropriate dialogue in morphologically rich languages like Korean.

Figure 16: Examples of common generation failures in Korean, illustrating two representative issues: (1) lexical redundancy in emotionally sensitive contexts, and (2) tone mismatches arising from inconsistent use of honorific and casual forms.

Example 1. Lexical Redundancy

深夜，小林站在母亲房门前犹豫了很久，终于鼓起勇气敲了门走了进去。
“妈妈，**我真的真的非常非常对不起您，我让您太失望了，真的真的不是故意的，我心里特别特别难受。**”

Late at night, Kobayashi stood in front of his mother’s door, hesitating for a long time before finally gathering the courage to knock and enter.
“Mom, I am really, really, so, so sorry. I’ve let you down so much. I truly didn’t mean to. I feel especially, especially awful inside.”

The example demonstrates emotional redundancy through repeated intensifiers (e.g., “真的真的”, “非常非常”, “特别特别”), which amplify the sentiment in an unnatural and overly dramatic manner. Such expressions are rarely used in authentic spoken Chinese and are stylistically misaligned with casual emotional disclosure. This overuse diminishes the perceived sincerity of the apology and undermines the dialogue’s pragmatic realism.

Example 2. Honorific Inconsistency

公司茶歇时，小王看见总监林林先生在清理咖啡机，连忙跑过去。“**林哥，您**今天辛苦了，我替您弄吧，**哥你**歇会儿。”

During the company coffee break, Xiao Wang saw Director Linlin cleaning the coffee machine and quickly ran over.
“Brother Lin, you’ve worked hard today. Let me handle this for you, take a rest.”

The example illustrates honorific inconsistency, where the speaker mixes formal and informal address terms within a single utterance (e.g., “林哥” with “**您**” and “**你**”). Such inconsistency in speech level creates pragmatic confusion, as the speaker simultaneously signals both deference and familiarity. This mismatch disrupts the perceived social relationship between interlocutors and results in stylistic incoherence, particularly problematic in languages like Chinese, where honorifics play a key role in encoding hierarchical and relational cues.

Example 3. Register Mixing

宿舍深夜，小陈碰见班主任李老师，想解释因为自己迟交了作业。
“李老师，关于那份作业**之延误，实乃**学生疏忽，**望您见谅，哪.....不好意思哈，真的不是故意的。**”

Late at night in the dormitory, Xiao Chen ran into his homeroom teacher, Mr. Li, and wanted to explain why he had submitted his assignment late.
“Mr. Li, the delay with that assignment was truly my oversight. I hope you can forgive me. I’m... really sorry, it honestly wasn’t intentional.”

The example reveals a register mismatch, where highly formal written expressions (e.g., “之延误”, “实乃”, “望您见谅”) are juxtaposed with informal colloquial phrases (e.g., “哪.....不好意思哈”, “真的不是故意的”) within the same utterance. This abrupt shift in style disrupts tonal consistency and weakens the pragmatic coherence of the dialogue. In Chinese, such mixing between classical written forms and casual spoken language is rarely observed in natural conversation. It often results in outputs that feel constructed or artificial rather than contextually grounded.

Figure 17: Examples of generation failures in Chinese, illustrating three common error types: (1) emotional redundancy from repeated intensifiers, (2) honorific inconsistency due to mixed formal and informal address terms, and (3) register mismatches from combining classical written expressions with colloquial speech.