HiCAP: Hierarchical Cultural Adaptation via Personality for Dialogue Generation

Lingzhi Shen[1], Xiaohao Cai[1], Yunfei Long[2], Imran Razzak[3], Shoaib Jameel[1]

[1]School of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom
[3]Department of Computational Biology, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

Abstract

Cultural adaptation in dialogue systems refers to the ability to generate responses that conform to linguistic norms and social expectations across different cultures. However, existing dialogue generation approaches pay limited attention to the role of personality in shaping cultural expression and communication style. We therefore propose HiCAP, a novel framework that integrates personality traits into culturally adaptive dialogue generation. HiCAP employs a personality-conditioned control mechanism that hierarchically modulates response style by first aligning it with the cultural tone and then refining linguistic expression and content according to individual personality attributes. To embed these factors, the model combines supervised fine-tuning and instruction-based tuning, supported by dynamic instruction templates and multi-agent demonstrations to capture diverse combinations of culture and personality. Experiments on CulturalPersonas and CPED show that HiCAP achieves the strongest overall performance across various dialogue generation metrics, demonstrating effective cultural adaptability and personality modelling.

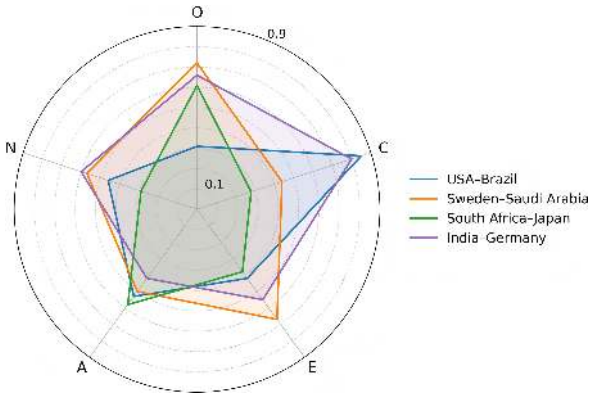


Figure 1: Pairwise Wasserstein distances for Big Five personality traits across four country pairs, with larger values indicating greater cross-cultural differences in trait distributions.

1 Introduction

The ability to understand and adapt to cultural context is a key step towards building human-centric artificial intelligence (AI) (Shen et al., 2026). In conversational agents, cultural adaptation (Anik et al., 2025) refers to the ability to flexibly adjust language and interaction strategies in accordance with the values, communication styles, and expectations associated with diverse cultural backgrounds, such as different countries, regions, or social groups. Such adaptability is essential for fostering trust and effective communication across global user populations. Accordingly, culturally aware dialogue agents have broad applications (Liu et al., 2025a), including international customer service, multilingual virtual assistants, cross-cultural healthcare support, global education platforms, and mental health counselling, where contextually appropriate and respectful interactions are particularly important for minority groups.

Recent research has advanced the cultural awareness of large language models (LLMs) through the use of novel data resources and targeted training strategies. Culture-Gen (Li et al., 2024d) proposes a prompting framework that systematically elicits cultural perceptions across countries and topics, enabling large-scale analysis of implicit cultural knowledge and bias without external knowledge bases or manual annotation. By contrast, CultureBank (Shi et al., 2024) constructs an open, structured knowledge base from culture-related narratives in online communities, using a pipeline that integrates automated filtering, LLM-assisted extraction, clustering, and human validation to capture fine-grained, community-driven cultural practices. CultureLLM (Li et al., 2024a) further improves scalability by leveraging representative questions from the World Value Survey (Haerpfer et al., 2020) and semantic data augmentation to efficiently expand training data for low-cost fine-tuning of both

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 885–899
November 6-10, 2026 ©2026 Association for Computational Linguistics

culture-specific and unified models.

Most existing approaches, however, tend to neglect the role of personality traits in shaping cross-cultural interactions (Pawar et al., 2025). At the behavioural level, culture and personality are closely intertwined, with cultural context influencing both the expression and distribution of personality traits across groups. Cross-cultural studies (McCrae and Terracciano, 2005; Schmitt et al., 2007) have shown that average levels of Big Five[1] (Costa and McCrae, 2008) traits vary substantially across countries and cultural contexts, a pattern further illustrated by the IPIP-NEO data (Johnson, 2014) in Figure 1. Personality is also systematically reflected in linguistic style (Yarkoni, 2010; Shen et al., 2025c; Chen et al., 2025). For example, high openness is associated with greater curiosity and the use of abstract or imaginative vocabulary, as reflected in words such as "poet" or "universe", whereas high conscientiousness favours structured, goal-oriented expressions and avoids overly casual language. For conversational agents, incorporating personality traits therefore provides an effective means of modelling cultural and social context.

In this paper, we propose **HiCAP** (**Hi**erarchical **C**ultural **A**daptation via **P**ersonality), which explicitly integrates personality traits into culturally adaptive dialogue generation. The core of HiCAP lies in a personality-conditioned control mechanism that first establishes a culture aligned baseline for tone and politeness, and then refines linguistic style and content selection according to individual personality traits, thereby enabling hierarchical modulation of response style. Specifically, during training, supervised fine-tuning (Dong et al., 2023) on dialogue data annotated with cultural and personality attributes allows the model to learn mappings between persona representations and stylistic adjustments through a gating mechanism. In addition, HiCAP employs slot-filling techniques (Mirza et al., 2024) to dynamically construct instruction templates, which are adaptively adjusted based on users' cultural background and personality traits. When explicit user persona information is unavailable, the framework infers the most likely cultural background and personality traits from the input, ensuring personalised instructions across different personality-culture combinations. Inspired by CulturePark (Li et al., 2024b), HiCAP further adopts multi-agent interaction to generate cross-cultural dialogue demonstrations, providing representative exemplars that illustrate the intended communicative style.

Our main contributions are as follows:

- We propose HiCAP, a framework that integrates personality traits into culturally adaptive dialogue generation through hierarchical style modulation, allowing cultural context to guide overall response style while personality traits refine individual expression.
- We develop persona-aware learning and instruction mechanisms that combine personality-conditioned gating, dynamic instruction construction, persona inference, and multi-agent demonstrations to support dialogue generation aligned with cultural context and personality traits.
- Experiments on CulturalPersonas and CPED demonstrate that HiCAP outperforms state-of-the-art methods across various dialogue generation metrics. Further analyses validate its cultural adaptability, personality modelling, and linguistic quality.

2 Related Work

Training-based Methods: This category comprises three main paradigms that include pre-training from scratch on culturally rich corpora, continued pre-training on culture-specific data, and supervised fine-tuning with culturally annotated datasets. Models such as HyperClovaX (Yoo et al., 2024) and PersianLLaMA (Abbasi et al., 2023) follow the first paradigm, achieving strong native-level cultural competence through large-scale region-specific training. To improve scalability, continued pre-training has been widely adopted to adapt general-purpose models to cultural domains, as exemplified by TaiwanLLM (Lin and Chen, 2023), Komodo (Owen et al., 2024), SeaLLM (Nguyen et al., 2023), and AceGPT (Huang et al., 2023). In parallel, supervised fine-tuning leverages labelled data capturing cultural, demographic, or social nuances, including Chinese culture-specific corpora (Bai et al., 2024), the Indonesian Cendol dataset (Cahyawijaya et al., 2024), and value survey-based resources for hate speech detection (Zhou et al., 2023). However, These

[1]The Big Five, also known as the Five-Factor Model, is the most widely accepted framework in personality psychology, describing five fundamental dimensions of human personality: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN).

training-based approaches generally incur substantial computational cost (Shen et al., 2025b).

Inference-time Methods: These methods adjust the generation process through prompt engineering that explicitly specifies cultural background, social roles, or value systems, without modifying model parameters. Representative strategies include culture-specific tokens or templates (Rao et al., 2024) and the explicit injection of cultural identity into prompts (Wang et al., 2023). Chain-of-thought and value-oriented prompting further encourage reasoning within diverse moral or cultural frameworks (Johnson et al., 2022; Masoud et al., 2023). Recent studies have also shown that language-specific prompt designs can mitigate bias towards dominant cultures, particularly for underrepresented groups (Naous et al., 2023). In addition, prompt-based cultural alignment tests and context-aware evaluation prompts have been used to reveal implicit cultural preferences and assess cultural adaptability across tasks (Cao et al., 2023; Fung et al., 2024). However, their capacity to support adaptation remains constrained when the underlying model lacks sufficient culture-specific knowledge.

HiCAP differs from both categories by jointly modelling culture and personality across training and inference. Persona-aware supervised fine-tuning embeds cultural and personality signals into the model's internal representations, while dynamic instructions and multi-agent demonstrations provide explicit, context-sensitive guidance during generation. A hierarchical personality-conditioned control mechanism further modulates response style by establishing a cultural baseline and refining it according to individual personality traits.

3 HiCAP Framework

As shown in Figure 2, HiCAP integrates personality traits into culturally adaptive dialogue generation through hierarchical personality-conditioned control. It combines persona-aware supervised fine-tuning (SFT), dynamic instruction construction, and multi-agent demonstrations. A hierarchical control mechanism uses cultural information to establish a shared stylistic baseline and personality information to refine expression at the token level.

3.1 Persona-Aware Learning

Persona-aware learning represents cultural background and personality traits as conditioning signals during SFT, allowing persona information to influence the model's internal representations.

We represent each persona using a structured profile that combines a cultural category with binary indicators for the Big Five personality dimensions. Let c denote the cultural category and $b_i \in \{0, 1\}$ denote the state of the i-th personality trait, where $b_i = 1$ and $b_i = 0$ correspond to high and low levels, respectively. The persona profile is $\mathbf{z} = [c, b_1, b_2, \ldots, b_{|\mathcal{P}|}]$, where $\mathcal{P}$ denotes the set of personality dimensions. Each discrete attribute is mapped to a trainable embedding, yielding a culture embedding $\mathbf{e}_c$ and personality embeddings $\mathbf{e}_{b_i}$. These embeddings are concatenated and projected into the hidden space of the backbone as $\mathbf{p} = \mathbf{W}_{\text{proj}} \,\text{Concat}(\mathbf{e}_c, \mathbf{e}_{b_1}, \ldots, \mathbf{e}_{b_{|\mathcal{P}|}})$, where $\mathbf{W}_{\text{proj}}$ is a learnable projection matrix and $\mathbf{p} \in \mathbb{R}^{d_h}$ is the resulting persona representation with hidden dimension d_h. The projected representation is used for dynamic instruction construction, while the culture and personality embeddings are retained for hierarchical modulation.

During SFT, let X denote the dialogue context and $Y = (y_1, \ldots, y_{|Y|})$ the target response. We optimise the standard next-token prediction loss $\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{|Y|} \log P_\Theta(y_t \mid y_{<t}, X, \mathbf{z})$, where Θ denotes the trainable model parameters.

To encourage the hidden representations to retain persona information, we additionally predict the cultural category and personality traits from a pooled context representation $\mathbf{h}_X$. The culture head produces $\hat{\mathbf{y}}_c = \text{softmax}(\mathbf{W}_{\text{cls},c}\mathbf{h}_X + \boldsymbol{\beta}_{\text{cls},c})$, where $\mathbf{W}_{\text{cls},c}$ and $\boldsymbol{\beta}_{\text{cls},c}$ are its trainable weight matrix and bias vector. For the i-th personality dimension, the corresponding prediction head produces $\hat{q}_i = \sigma(\mathbf{w}_{\text{cls},i}^\top \mathbf{h}_X + \beta_{\text{cls},i})$, where $\mathbf{w}_{\text{cls},i}$ and $\beta_{\text{cls},i}$ are trainable parameters and $\sigma(\cdot)$ denotes the sigmoid function. We denote the resulting auxiliary persona classification loss by $\mathcal{L}_{\text{cls}} = \ell_{\text{cls}}(\mathbf{z}, \mathbf{h}_X)$, where ℓ_{cls} jointly measures errors in cultural-category and personality-trait prediction. This auxiliary supervision encourages the internal representation to remain aligned with the intended persona during training.

3.2 Dynamic Instruction Construction

Dynamic instruction construction converts the structured persona representation into explicit natural-language guidance for response generation. Let $\mathcal{T}$ denote a curated set of instruction template schemas with parameterised slots. A tem-

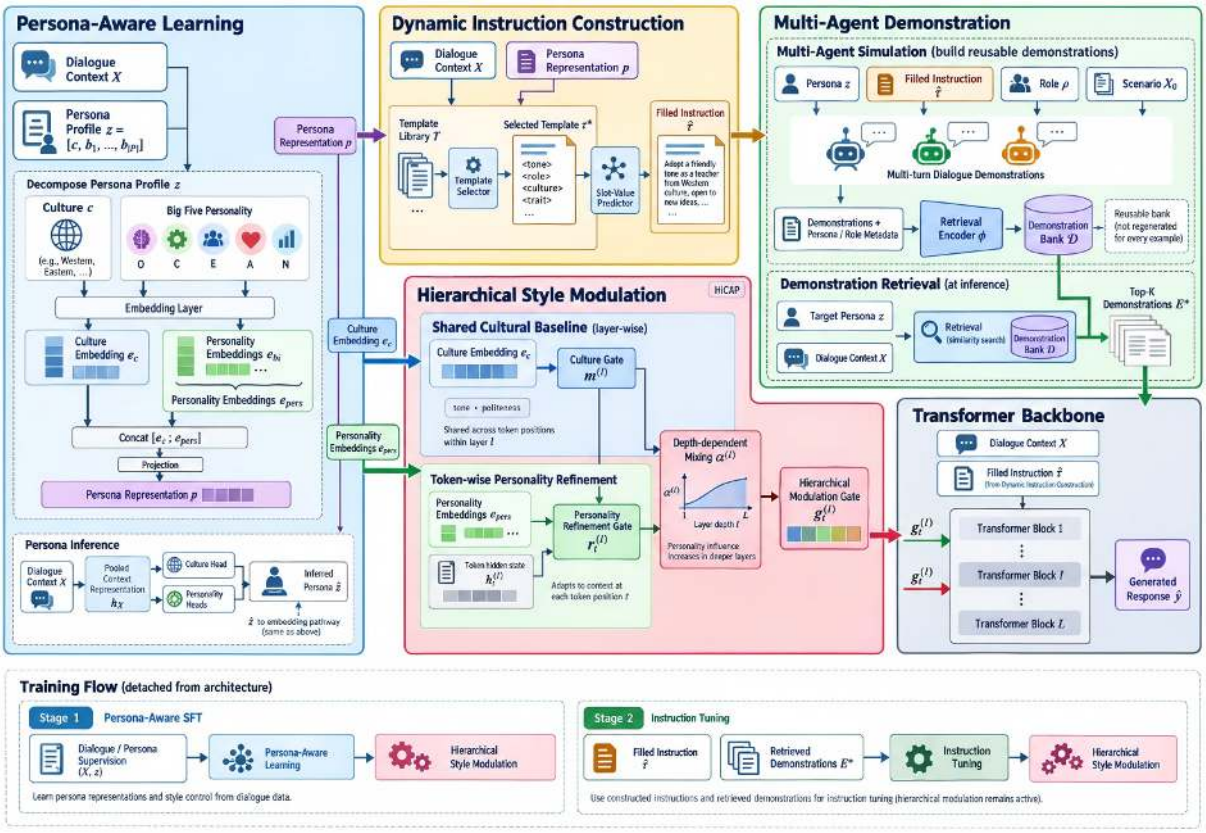


Figure 2: Overview of HiCAP, combining persona-aware learning, dynamic instruction construction, multi-agent demonstrations, and hierarchical style modulation to generate culturally and personality-aligned dialogue.

plate $\tau \in \mathcal{T}$ is a sequence of ordinary tokens interleaved with semantic placeholders, such as ⟨tone⟩ or ⟨role⟩. Formally, $\tau = (t_1, \ldots, \langle \text{slot}_k \rangle, \ldots, t_{n_\tau})$, where n_τ is the number of elements in the template and k indexes a slot. Each slot specifies a controllable attribute related to culture, personality, or interaction context. For the k-th slot, a candidate set $\mathcal{V}_k$ is constructed from high-frequency expressions in labelled dialogue corpora and manually curated variants to maintain stylistic coverage and linguistic naturalness.

Table 5 (Appendix A) summarises the slot categories, slot types, and representative values used in the dynamic templates.

For a persona representation $\mathbf{p}$ and dialogue context X, each template receives a compatibility score $s_{\text{temp}}(\tau \mid \mathbf{p}, X)$ from a learned scoring function. The highest-scoring template is selected as $\tau^* = \arg\max_{\tau \in \mathcal{T}} s_{\text{temp}}(\tau \mid \mathbf{p}, X)$. Each slot in τ^* is then instantiated with a context-sensitive value $v_k \in \mathcal{V}_k$ selected by a learned slot-value predictor. The resulting instruction $\hat{\tau} = \text{FillSlots}(\tau^*, \{v_k\}_k)$ expresses the relevant cultural, personality, and interaction guidance in natural language and is used as explicit conditioning during instruction tuning and generation.

When explicit persona annotations are unavailable, HiCAP infers the persona profile from the dialogue context. Using the prediction heads defined in Section 3.1, the cultural category is estimated as $\hat{c} = \arg\max_j \hat{y}_{c,j}$, where $\hat{y}_{c,j}$ is the predicted probability of cultural category j. Each personality indicator is obtained as $\hat{b}_i = \mathbb{I}[\hat{q}_i \geq \gamma]$, where γ is a fixed decision threshold and $\mathbb{I}[\cdot]$ is the indicator function. The inferred profile $\hat{\mathbf{z}} = [\hat{c}, \hat{b}_1, \ldots, \hat{b}_{|\mathcal{P}|}]$ is mapped to the persona representation $\mathbf{p}$ using the same embedding and projection functions as an annotated profile.

3.3 Multi-Agent Demonstration

Inspired by CulturePark (Li et al., 2024b), we use automated multi-agent simulation to generate culturally grounded and personality-consistent dialogue demonstrations. Each simulated agent is assigned a structured persona profile $\mathbf{z}$ and its corresponding dynamic instruction $\hat{\tau}$, allowing cultural background and personality traits to shape the simulated interaction.

Let $\rho \in \mathcal{V}_{\text{role}}$ denote an interaction role, such as `advisor` or `friend`, where $\mathcal{V}_{\text{role}}$ is the set of available roles. Given a scenario seed X_0, which provides a brief description of the interaction setting, the relevant role is instantiated in the corresponding instruction slot. The initial prompt for an agent is then constructed as $\Pi(\mathbf{z}, \rho, X_0) = [\,\hat{\tau} \oplus C_{\text{sys}} \oplus X_0\,]$, where C_{sys} denotes the system context and $\oplus$ denotes sequence concatenation.

During simulation, agents take turns generating utterances under their respective persona configurations, producing multi-turn dialogue trajectories across different cultural backgrounds, personality traits, roles, and scenarios. Persona conditioning remains active during generation, allowing these attributes to influence lexical choice and stylistic realisation throughout the interaction.

The generated demonstrations are stored with metadata describing their persona attributes, roles, and interaction contexts. For each demonstration u, let $\mathbf{a}_u = [c, b_1, \ldots, b_{|\mathcal{P}|}, \rho]$ denote its structured metadata. Let $\phi(\cdot)$ denote a retrieval encoder that maps metadata and text into a shared semantic space. The demonstration is encoded as $\mathbf{e}_u = \phi(\mathbf{a}_u \oplus u)$.

Given a target persona–context pair $(\mathbf{z}^*, X^*)$, the query is encoded as $\mathbf{e}_q = \phi(\mathbf{z}^* \oplus X^*)$ in the same semantic space. For each demonstration u in the demonstration bank $\mathcal{D}$, the retrieval score is $s_{\text{ret}}(u \mid \mathbf{z}^*, X^*) = \text{sim}(\mathbf{e}_u, \mathbf{e}_q)$, where $\text{sim}(\cdot, \cdot)$ denotes similarity in the embedding space. The K highest-scoring demonstrations, where K is the number of retrieved exemplars, form the set $\mathcal{E}^*$.

After persona-aware SFT, HiCAP performs instruction tuning with the filled instruction and retrieved demonstrations as additional conditioning. For a training pair $(X, Y, \mathbf{z})$, instruction tuning optimises the standard next-token loss $\mathcal{L}_{\text{IT}} = -\sum_{t=1}^{|Y|} \log P_\Theta(y_t \mid y_{<t}, X, \mathbf{z}, \hat{\tau}, \mathcal{E}^*)$. At inference time, the same retrieval procedure supplies $\mathcal{E}^*$ as culturally and personality-aligned few-shot exemplars for response generation.

3.4 Hierarchical Style Modulation

Hierarchical style modulation controls how cultural and personality information influence generation. Cultural information establishes a shared stylistic baseline, including properties such as tone and politeness, while personality traits refine expression according to the current token context.

Let L denote the number of transformer blocks and T the number of token positions. We retain the culture embedding $\mathbf{e}_c$ and combine the personality embeddings as $\mathbf{e}_{\text{pers}} = \text{Concat}(\mathbf{e}_{b_1}, \ldots, \mathbf{e}_{b_{|\mathcal{P}|}})$. At transformer block $l \in \{1, \ldots, L\}$, the culture gate $\mathbf{m}^{(l)} = \sigma(F^{(l)}_{\text{cult}}(\mathbf{e}_c))$ defines a baseline shared across token positions. For token position $t \in \{1, \ldots, T\}$, the personality refinement gate $\mathbf{r}^{(l)}_t = \sigma(F^{(l)}_{\text{pers}}([\mathbf{h}^{(l)}_t \| \mathbf{e}_{\text{pers}}]))$ adapts the modulation to the evolving hidden state $\mathbf{h}^{(l)}_t$, where $F^{(l)}_{\text{cult}}$ and $F^{(l)}_{\text{pers}}$ are trainable gate networks and $\|$ denotes vector concatenation.

We use a learned depth-dependent mixing coefficient $\alpha^{(l)} \in (0, 1)$ that increases with layer depth, giving personality refinement progressively greater influence in deeper transformer blocks. The final token-wise modulation gate is

$$\mathbf{g}^{(l)}_t = \big(1 - \alpha^{(l)}\big)\mathbf{m}^{(l)} + \alpha^{(l)}\mathbf{r}^{(l)}_t. \tag{1}$$

Using the gate in Equation 1, the hidden state is updated as

$$\mathbf{h}^{(l+1)}_t = \mathbf{g}^{(l)}_t \odot f^{(l)}_\Theta\Big(\text{LN}(\mathbf{h}^{(l)}_t)\Big) + \big(\mathbf{1} - \mathbf{g}^{(l)}_t\big) \odot \mathbf{h}^{(l)}_t, \tag{2}$$

where $f^{(l)}_\Theta(\cdot)$ denotes the transformation performed by block l, $\text{LN}(\cdot)$ denotes layer normalisation, $\odot$ denotes element-wise multiplication, and $\mathbf{1}$ is an all-ones vector with the same dimensionality as the gate. The residual term preserves the contextual representation, while persona dependence enters through the gate controlling the transformed and residual states.

To keep personality refinement anchored to the cultural baseline, let $\bar{\mathbf{g}}^{(l)} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{g}^{(l)}_t$ denote the gate averaged across token positions at layer l. We define the hierarchical regulariser as

$$\mathcal{L}_{\text{hier}} = \sum_{l=1}^{L} \Big\| \bar{\mathbf{g}}^{(l)} - \mathbf{m}^{(l)} \Big\|_2^2. \tag{3}$$

Equation 3 penalises the average deviation of the token-wise modulation gate from the cultural baseline while retaining token-specific variation. Let $\mathcal{L}_{\text{gen}}$ denote $\mathcal{L}_{\text{SFT}}$ during persona-aware SFT and $\mathcal{L}_{\text{IT}}$ during instruction tuning. Combining the corresponding next-token loss, auxiliary persona supervision, and hierarchical regularisation, the training objective for each stage is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gen}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{hier}}\mathcal{L}_{\text{hier}}, \tag{4}$$

Algorithm 1 HiCAP Training and Inference

Require: Context X, persona $\mathbf{z}$, optional target Y, demonstration bank $\mathcal{D}$, parameters Θ, learning rate η

if Y is available **then**
  Compute $\mathcal{L}_{\text{total}}$ on $(X, Y, \mathbf{z})$ with layer-wise hierarchical modulation and $\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{SFT}}$
  Update Θ using $\mathcal{L}_{\text{total}}$ with learning rate η
  $(\mathbf{p}, \mathbf{e}_c, \mathbf{e}_{\text{pers}}) \leftarrow \textsc{Embed}(\mathbf{z})$
  $\tau^* \leftarrow \textsc{Template}(\mathbf{p}, X)$; $\hat{\tau} \leftarrow \textsc{Fill}(\tau^*, \mathbf{p}, X)$
  $\mathcal{E}^* \leftarrow \textsc{Retrieve}(\mathcal{D}, \mathbf{z}, X)$
  Compute $\mathcal{L}_{\text{total}}$ on $(X, \hat{\tau}, \mathcal{E}^*, Y, \mathbf{z})$ with layer-wise hierarchical modulation and $\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{IT}}$
  Update Θ using $\mathcal{L}_{\text{total}}$ with learning rate η
else
  Infer $\mathbf{z} \leftarrow \textsc{Infer}(X)$ if unavailable
  $(\mathbf{p}, \mathbf{e}_c, \mathbf{e}_{\text{pers}}) \leftarrow \textsc{Embed}(\mathbf{z})$
  $\tau^* \leftarrow \textsc{Template}(\mathbf{p}, X)$; $\hat{\tau} \leftarrow \textsc{Fill}(\tau^*, \mathbf{p}, X)$
  $\mathcal{E}^* \leftarrow \textsc{Retrieve}(\mathcal{D}, \mathbf{z}, X)$
  $\hat{Y} \leftarrow \textsc{Decode}(X, \hat{\tau}, \mathcal{E}^*)$ with layer-wise hierarchical modulation
  return $\hat{Y}$
end if

where λ_{cls} and λ_{hier} control the contributions of persona supervision and hierarchical regularisation, respectively.

Algorithm 1 summarises how the components described above are integrated in the training and inference processes of HiCAP.

4 Experiments and Results

4.1 Experimental Setup

Datasets: We adopt CulturalPersonas (Dey et al., 2025) to evaluate personality expression across different cultural contexts and CPED (Chen et al., 2022) to examine personality expression and cultural adaptation across different social groups within a single society. Together, they provide complementary cross-cultural and intra-cultural evaluation settings. CulturalPersonas is a large-scale, human-validated benchmark designed to assess how LLMs express personality traits across authentic cultural contexts. It consists of 3,000 scenario-based question–answer pairs from six countries (United States, Brazil, India, Japan, Saudi Arabia, and South Africa). Each instance is associated with one of the Big Five personality dimensions and supports both Likert-style multiple-choice and open-ended evaluation. CPED is a large-scale Chinese dialogue dataset containing over 133,000 utterances across more than 12,000 multi-turn dialogues extracted from 40 contemporary Chinese TV dramas. Utterances are annotated with sentiment, emotion, dialogue acts, scene categories, and static personal attributes, including gender, age, and Big Five traits.

Implementation Details: We use the BPDD dataset (Liu et al., 2025b) for supervised fine-tuning and Phi-4-14B (Abdin et al., 2024) as the backbone. The model is trained in bfloat16 precision with a maximum context length of 16K tokens on an NVIDIA H200 cluster. Both training stages use the AdamW optimiser with a cosine learning-rate schedule. Newly introduced modules use a learning rate of 2×10^{-5}, while the backbone uses 5×10^{-6}. The dynamic instruction library contains 36 template schemas with 4–9 parameterised slots each. Culture and personality embeddings have dimensions of 128 and 64, respectively. The culture and personality gate networks use two-layer MLPs with hidden size 2048, while the slot-value predictor uses an MLP with hidden size 512; all newly introduced MLPs use GELU activation. The demonstration index contains approximately 1.2 million persona-aligned utterances, with synthetic demonstrations capped at 30% of the conditioning tokens during instruction tuning to mitigate overfitting. Decoding uses nucleus sampling with $p = 0.95$ and temperature 0.7.

Evaluation Metrics: Following the evaluation settings of CulturalPersonas and CPED, we adopt several complementary evaluation metrics, including BLEU for quantifying surface-level n-gram overlap, BERTScore for measuring semantic similarity between model outputs and reference texts using contextual embeddings, Perplexity for assessing generation fluency and predictive confidence, and the type–token ratio (TTR) together with Distinct-n for capturing vocabulary richness and variability, where TTR focuses on the unique word ratio and Distinct-n extends this to n-gram diversity. These metrics are jointly well suited for cross-cultural evaluation, as cultural differences are often reflected in variations in lexical choice, semantic framing, and stylistic diversity.

Comparative Models: For CPED, we include Seq2Seq (Sutskever et al., 2014), Transformer

Dataset	Model	BLEU ↑	BERTScore ↑	Perplexity ↓	TTR ↑	Distinct-1 ↑	Distinct-2 ↑
CPED	Seq2Seq	0.0077	0.5196	107.30	0.4342	0.0252	0.1846
	Transformer	0.1680	0.5519	62.82	0.5687	0.0264	0.2031
	CDial-GPT	0.1403	0.5719	17.70	0.6412	0.0602	0.3388
	Qwen-2.5-72B-Instruct	0.1710	0.6815	15.50	0.6910	0.0707	0.3587
	Llama-3.3-70B-Instruct	0.1621	0.6705	15.92	0.6820	0.0663	0.3451
	Gemma-2-27B-IT	0.1553	0.6642	16.88	0.6755	0.0646	0.3385
	DeepSeek-V3.1	0.1748	0.6889	15.15	0.6928	0.0723	0.3640
	GPT-4o	0.1821	0.7035	14.88	0.6994	0.0752	0.3743
	HiCAP (Ours)	**0.1884**	**0.7112**	**14.72**	**0.7028**	**0.0761**	**0.3812**
CulturalPersonas	Qwen-2.5-72B-Instruct	0.1639	0.6764	16.38	0.7162	0.0773	0.3776
	Llama-3.3-70B-Instruct	0.1612	0.6727	16.22	0.7141	0.0763	0.3731
	Gemma-2-27B-IT	0.1576	0.6683	16.66	0.7083	0.0742	0.3671
	Kimi-K2-Instruct	0.1648	0.6849	16.03	0.7256	0.0823	0.3909
	DeepSeek-V3.1	0.1677	0.6814	15.88	0.7223	0.0804	0.3862
	GPT-4o	0.1714	0.6936	**15.62**	0.7281	0.0842	**0.4023**
	HiCAP (Ours)	**0.1762**	**0.7001**	15.78	**0.7309**	**0.0853**	0.3919

Table 1: Dialogue generation performance of different models on CPED and CulturalPersonas. ↑ indicates higher is better; ↓ indicates lower is better.

(Vaswani et al., 2017), and CDial-GPT (Wang et al., 2020) models fine-tuned on the dataset. We further compare HiCAP with several advanced LLMs, including Qwen-2.5 (Team, 2024), Llama-3.3 (Dubey et al., 2024), Gemma-2 (Team et al., 2024), DeepSeek-V3.1 (Liu et al., 2024a), and GPT-4o (Hurst et al., 2024). Kimi-K2 (Team et al., 2025) is additionally included for CulturalPersonas.

4.2 Overall Results

Table 1 reports dialogue generation performance on CPED and CulturalPersonas. On CPED, HiCAP achieves the best performance across all six reported metrics. On CulturalPersonas, HiCAP achieves the highest BLEU, BERTScore, TTR, and Distinct-1, while GPT-4o obtains lower Perplexity and higher Distinct-2. The consistency of the gains across two substantially different dialogue settings suggests that HiCAP remains effective under both Chinese multi-turn dialogue and cross-cultural scenario-based generation.

The metric-level results reveal several distinct patterns. Improvements in both BLEU and BERTScore indicate that HiCAP increases agreement with reference responses at both the lexical and semantic levels, with BERTScore gains showing that the improvement is not confined to surface-form overlap. On CPED, HiCAP also achieves the lowest Perplexity while simultaneously obtaining the highest TTR and Distinct scores, indicating that greater lexical diversity is achieved without an accompanying degradation in sequence likelihood. CulturalPersonas presents a slightly different pattern. HiCAP attains the highest TTR and Distinct-1, suggesting broader lexical usage, whereas GPT-4o achieves the highest Distinct-2 and slightly lower Perplexity. This difference suggests that HiCAP favours vocabulary-level diversity, while GPT-4o exhibits greater diversity in local bigram combinations on this benchmark.

The comparison across model families further highlights the relative strength of HiCAP. Conventional dialogue models, including Seq2Seq, Transformer, and CDial-GPT, remain substantially behind recent LLMs, particularly in semantic similarity and lexical diversity. Among the general-purpose LLMs, GPT-4o remains the strongest baseline on most metrics across the two benchmarks, with models such as DeepSeek-V3.1 following closely, while Kimi-K2 also demonstrates strong lexical diversity. Despite these competitive baselines, HiCAP achieves the strongest overall performance across both benchmarks. Notably, HiCAP uses a 14B backbone while competing with substantially larger models ranging from 27 billion to 1 trillion parameters, suggesting that the observed gains cannot be explained by model scale alone.

4.3 Ablation Study

We conduct an ablation study of the key components of HiCAP, as shown in Table 2. We focus on three complementary generation metrics: Distinct-1 (D-1), Distinct-2 (D-2), and Perplexity (PPL), which jointly assess lexical diversity and fluency.

We first examine the impact of replacing the backbone language model. Using the vanilla Phi-4 model without HiCAP components yields the lowest diversity and the highest perplexity, indicating the contribution of the proposed components to

Component	D-1 ↑	D-2 ↑	PPL ↓
Vanilla Phi-4	0.0697	0.3228	18.04
Replace Phi-4 with Qwen-2.5-14B	0.0850	0.3913	**15.64**
Replace Phi-4 with Llama-3.1-8B	0.0823	0.3754	16.29
w/o Persona-Conditioned Gating	0.0729	0.3397	17.04
Replace Token-Wise with Scalar	0.0761	0.3538	16.93
w/o Personality Inference	0.0784	0.3602	16.47
w/o Dynamic Template Selector	0.0738	0.3441	16.52
w/o Demonstrations	0.0792	0.3653	15.98
w/o Instruction Tuning	0.0733	0.3426	17.18
w/o Supervised Fine-Tuning	0.0764	0.3544	16.91
HiCAP	**0.0853**	**0.3919**	15.78

Table 2: Ablation results for HiCAP on CulturalPersonas. ↑ indicates higher is better; ↓ indicates lower is better.

generation quality. Replacing the Phi-4 backbone with Qwen-2.5-14B achieves comparable diversity with slightly lower perplexity, while Llama-3.1-8B produces intermediate results. This indicates that the framework generalises across different backbones, while the stronger performance of the larger models may also reflect known scaling effects (Kaplan et al., 2020).

The removal of persona-conditioned gating causes a clear drop in both diversity metrics and an increase in perplexity, highlighting its central role in integrating personality and cultural features into the backbone model. Without this mechanism, the stylistic control learned from these signals is largely lost. Replacing token-wise gating with a scalar alternative likewise reduces diversity and increases perplexity, although the degradation is less severe than when gating is entirely absent. This comparison confirms that fine-grained, token-level control is more effective than a single global coefficient for modulating the influence of personality and cultural information during generation.

Eliminating the personality inference branch results in moderate reductions in both diversity and fluency. Even when cultural information is preserved, the lack of personality predictions limits the model's ability to vary stylistic and lexical choices, particularly when explicit personality labels are unavailable in the input. Disabling the dynamic template selector leads to a more pronounced drop in diversity, as the model is constrained to a single static template with fixed slot values, reducing its adaptability across contexts. By contrast, demonstrations exert a relatively limited impact on generation quality. Although they facilitate alignment with context-specific styles, the backbone retains strong generation capability in their absence.

The final ablations examine the two post-training stages of HiCAP. Removing instruction tuning leads to a clear decrease in both diversity metrics and an increase in perplexity, underscoring its contribution to context-sensitive generation while maintaining fluency. Persona-aware supervised fine-tuning also contributes substantially, although its removal results in a smaller degradation across all three metrics. This pattern suggests that both post-training stages are important, with instruction tuning contributing more strongly to the measured diversity and fluency of the final model.

4.4 Additional Analyses

4.4.1 Implicit Personality Recognition

To evaluate the personality inference capability of HiCAP when explicit persona annotations are unavailable, we assess its ability to recover the Big Five traits directly from dialogue context on the CPED dataset. In this setting, recognition depends on subtle linguistic cues, pragmatic tendencies, and context-dependent expressions because explicit trait descriptions are absent (Shen et al., 2025a).

Methods	AGR	CON	EXT	NEU	OPN	Avg	F1
SVM	70.83	57.90	60.89	47.76	49.47	57.37	54.74
LSTM	70.01	60.56	64.73	53.75	54.93	60.80	59.81
BERT	75.89	63.48	68.21	52.67	55.42	63.13	61.47
RoBERTa	78.08	63.72	67.65	53.60	55.58	63.73	61.98
TrigNet	82.76	66.52	69.15	58.52	54.50	66.29	63.45
PS-GCN	82.87	67.75	72.48	58.50	56.60	67.64	64.80
GPT-4o	80.16	65.40	71.59	59.80	53.29	66.05	63.54
PsyCoT	80.18	65.83	69.43	53.35	56.15	64.99	62.77
EERPD	84.03	73.91	74.29	52.67	56.44	68.27	67.06
HiCAP	**85.36**	**75.18**	**74.92**	**60.24**	**57.31**	**70.60**	**68.74**

Table 3: Performance comparison (%) for Big Five personality recognition on CPED.

For this task, we compare against classic machine learning and neural models such as SVM (Tighe et al., 2016), LSTM (Tandera et al., 2017), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019), together with methods developed specifically for Big Five personality detection, including TrigNet (Yang et al., 2021), PS-GCN (Liu et al., 2024b), PsyCoT (Yang et al., 2023), and EERPD (Li et al., 2024e). As shown in Table 3, HiCAP achieves the highest average accuracy and Macro-F1 among all compared methods, outperforming EERPD by 2.33% and 1.68%, respectively. The gains across all five traits suggest that the personality inference mechanism can reliably recover personality information from implicit linguistic and contextual signals even when explicit trait descriptions are unavailable.

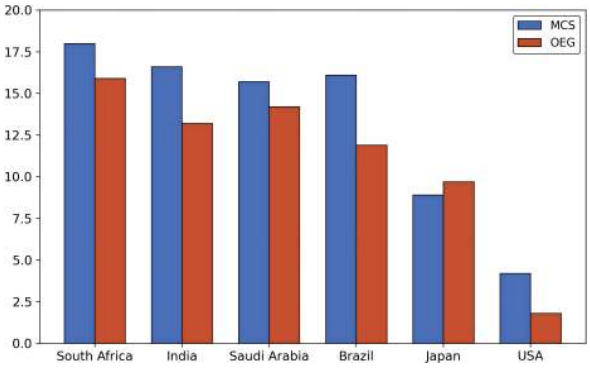


Figure 3: Cross-country comparison of HiCAP's improvement in cultural–personality alignment from cultural priming on CulturalPersonas.

4.4.2 Cultural Priming Effects

Cultural priming explicitly provides cultural cues at inference time, allowing the model to condition generation on a specified cultural frame. Although HiCAP already infers and encodes cultural information internally, explicit cultural cues may further reduce uncertainty in culturally conditioned expression.

As shown in Figure 3, cultural priming produces additional improvements under both the Likert-style multiple-choice (MCS) and open-ended generation (OEG) settings on CulturalPersonas, measured by relative reductions in Wasserstein distance. The largest marginal gains are observed in South Africa, India, and Saudi Arabia. This pattern may reflect greater divergence between these cultural contexts and dominant pretraining distributions, making explicit cultural grounding more beneficial. The gains in the United States are smaller, possibly because HiCAP's internal cultural inference is already closer to the target cultural priors without additional prompting. These results suggest that cultural priming provides greater benefit when the model's internal cultural representation is less closely aligned with the target context.

4.4.3 Automatic Linguistic Evaluation

To evaluate the linguistic quality of HiCAP's generated dialogues, we use an LLM-as-a-judge framework (Li et al., 2024c) on CulturalPersonas and CPED. The framework assesses fluency and coherence based on grammatical naturalness, semantic appropriateness, and contextual consistency within the dialogue. LLM-based evaluation has been widely used for assessing generated text and has shown strong agreement with human linguistic judgements (Zheng et al., 2023).

Evaluator	Fluency (%)	Coherence (%)
GPT-5	98.8	98.5
Claude-Sonnet-4.5	98.3	98.2
Gemini-2.5-Pro	98.1	97.8
Macro Average	**98.4**	**98.2**

Table 4: Evaluation of HiCAP's dialogue quality by three external LLM-based evaluators.

Specifically, three external evaluators, GPT-5 (Singh et al., 2025), Claude-Sonnet-4.5[2], and Gemini-2.5-Pro (Comanici et al., 2025), are employed to assess HiCAP's outputs using a unified evaluation prompt, provided in Appendix B and shown in Figure 4.

As shown in Table 4, HiCAP achieves average fluency and coherence scores of 98.4% and 98.2%, respectively. GPT-5 yields the highest scores, followed closely by Claude-Sonnet-4.5 and Gemini-2.5-Pro. The small differences across evaluators may reflect variations in their sensitivity to stylistic diversity and discourse continuity. Their closely aligned scores indicate consistent assessments of HiCAP's output quality. The slightly lower coherence scores may arise from conversational fillers and elliptical constructions that occur naturally in dialogue, while the high average scores indicate that the generated responses remain generally fluent and contextually coherent.

4.5 Case Analysis

Qualitative analyses of culturally adaptive response generation and safety risks are provided in Appendix C.1 and Appendix C.2, respectively.

5 Conclusions

We proposed HiCAP, a framework for culturally adaptive dialogue generation that uses personality traits to capture individual variation in cultural expression. HiCAP combines hierarchical personality-conditioned control with persona-aware learning, dynamic instruction construction, and multi-agent demonstrations. Experiments on CulturalPersonas and CPED demonstrate strong cultural adaptability, personality expressiveness, and overall generation quality, while ablation studies validate the contributions of the key components. Future work will explore finer-grained cultural and personality representations and extend HiCAP to multilingual and multimodal settings.

[2]Claude-Sonnet-4.5 System Card

6 Limitations

HiCAP has several limitations. Cultural awareness in dialogue systems remains dependent on the availability and diversity of training data. Current corpora are heavily skewed towards high-resource and English-centric settings, with limited coverage of low-resource languages, regional dialects, and subcultural contexts. Extending beyond text to culturally representative images and videos could further strengthen contextual grounding and understanding of local norms, helping to mitigate cultural bias and improve global applicability.

More refined cultural modelling is also required. Many existing approaches rely on coarse labels such as country or language, which are insufficient to capture intra-cultural variation. Integrating psychological personality models with anthropological frameworks may support richer cultural representations. Such dimensions should be theoretically grounded and empirically meaningful. Combining fine-grained cultural factors with personality traits may enable more precise and context-sensitive adaptation without oversimplification.

Another limitation concerns the dynamic nature of culture. Current language models have limited capacity to continuously update their cultural knowledge as linguistic norms and social practices evolve. Continual learning or online adaptation could enable models to absorb emerging patterns while preserving accumulated knowledge. Human–AI co-adaptation based on feedback from specific communities may further support evolving cultural competence.

Finally, personality modelling remains challenging in terms of stability and reliability. Minor prompt variations can lead to substantial shifts in inferred personas, and certain personality configurations may increase the likelihood of unsafe or impolite outputs. More robust personality control mechanisms are needed to maintain behavioural coherence across contexts while balancing expressive flexibility with safety constraints.

7 Ethical Considerations

Personality and culture are deeply interconnected yet highly variable across individuals and societies. Treating them as fixed or tightly coupled attributes risks reinforcing stereotypes and obscuring within-culture diversity. Apparent differences in self-reported traits may reflect social norms or measurement bias rather than genuine psychological variation. Hard-coding "typical personalities" for specific cultures can therefore produce homogeneous responses that fail to capture individual differences. Culturally and personally adaptive systems should model these factors probabilistically and contextually, preserving flexibility across individuals and situations.

The integration of cultural and personality signals raises concerns about bias and privacy. Persona conditioning may amplify social stereotypes or reproduce biases present in training data, such as associating gender or nationality with particular traits. Datasets annotated with cultural and personality information may also contain sensitive personal data. Responsible development therefore requires safeguards for data collection, anonymisation, and use, together with transparent evaluation procedures to identify and mitigate bias.

Personalisation and diversity must also be balanced with fairness and user autonomy. Techniques such as on-device learning and federated aggregation can support user-specific adaptation while reducing the risk of data exposure. Users should retain explicit control over interaction styles and personality settings, ensuring that personalisation remains transparent and consensual. Evaluation frameworks should also emphasise multi-perspective alignment, allowing dialogue agents to represent diverse cultural viewpoints without enforcing a single normative stance.

Acknowledgments

This work was supported by the Turing's Defence and Security programme through a partnership with the UK government in accordance with the framework agreement between GCHQ & The Alan Turing Institute.

References

Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei Bidgoli. 2023. Persianllama: Towards building first persian large language model. *arXiv preprint arXiv:2312.15713*.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Mahfuz Ahmed Anik, Abdur Rahman, Azmine Toushik Wasi, and Md Manjurul Ahsan. 2025. Pre-

serving cultural identity with context-aware translation through multi-agent ai systems. *arXiv preprint arXiv:2503.04827*.

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Junting Zhou, Ziqiang Liu, Feiteng Fang, Mingshan Chang, Tianyu Zheng, Xincheng Zhang, et al. 2024. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. *arXiv preprint arXiv:2403.18058*.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Afina Putri, Emmanuel Dave, Jhonson Lee, Nuur Shadieq, Wawan Cenggoro, Salsabil Maulana Akbar, Muhammad Ihza Mahendra, et al. 2024. Cendol: Open instruction-tuned generative large language models for indonesian languages. *arXiv preprint arXiv:2404.06138*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

Guanming Chen, Lingzhi Shen, Xiaohao Cai, Imran Razzak, and Shoaib Jameel. 2025. Hippd: Brain-inspired hierarchical information processing for personality detection. *arXiv preprint arXiv:2510.09893*.

Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. 2022. Cped: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *arXiv preprint arXiv:2205.14727*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Paul T Costa and Robert R McCrae. 2008. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2(2):179–198.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Priyanka Dey, Yugal Khanter, Aayush Bothra, Jieyu Zhao, and Emilio Ferrara. 2025. Can llms express personality across cultures? introducing culturalpersonas for evaluating trait alignment. *arXiv preprint arXiv:2506.05670*.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

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

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2020. World values survey wave 7 (2017-2020) cross-national data-set. *(No Title)*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, et al. 2023. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

John A Johnson. 2014. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. *Journal of research in personality*, 51:78–89.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37:65183–65216.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024c. Llms-as-judges: a comprehensive survey

on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Huihan Li, Liwei Jiang, Jena D Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024d. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *arXiv preprint arXiv:2404.10199*.

Zheng Li, Dawei Zhu, Qilong Ma, Weimin Xiong, and Sujian Li. 2024e. Eerpd: Leveraging emotion and emotion regulation for improving personality detection. *arXiv preprint arXiv:2406.16079*.

Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025a. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Wenjuan Liu, Zhengyan Sun, Subo Wei, Shunxiang Zhang, Guangli Zhu, and Lei Chen. 2024b. Ps-gcn: Psycholinguistic graph and sentiment semantic fused graph convolutional networks for personality detection. *Connection Science*, 36(1):2295820.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhi Liu, Yao Xiao, Zhu Su, Luyao Ye, Kaili Lu, and Xian Peng. 2025b. Bilingual dialogue dataset with personality and emotion annotations for personality recognition in education. *Scientific Data*, 12(1):514.

Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.

Robert R McCrae and Antonio Terracciano. 2005. Personality profiles of cultures: aggregate personality traits. *Journal of personality and social psychology*, 89(3):407.

Paramita Mirza, Viju Sudhi, Soumya Ranjan Sahoo, and Sinchana Ramakanth Bhat. 2024. Illuminer: Instruction-tuned large language models as few-shot intent classifier and slot filler. *arXiv preprint arXiv:2403.17536*.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, et al. 2023. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

Louis Owen, Vishesh Tripathi, Abhay Kumar, and Biddwan Ahmed. 2024. Komodo: a linguistic expedition into indonesia's regional languages. *arXiv preprint arXiv:2403.09362*.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *CoRR*.

David P Schmitt, Jüri Allik, Robert R McCrae, and Verónica Benet-Martínez. 2007. The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of cross-cultural psychology*, 38(2):173–212.

Lingzhi Shen, Xiaohao Cai, Yunfei Long, Imran Razzak, Guanming Chen, and Shoaib Jameel. 2025a. Emoperso: Enhancing personality detection with self-supervised emotion-aware modelling. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 2577–2587.

Lingzhi Shen, Xiaohao Cai, Yunfei Long, Imran Razzak, Guanming Chen, and Shoaib Jameel. 2026. Calm: Culturally self-aware language models. *Advances in Neural Information Processing Systems*, 38:118061–118094.

Lingzhi Shen, Yunfei Long, Xiaohao Cai, Guanming Chen, Imran Razzak, and Shoaib Jameel. 2025b. Less but better: Parameter-efficient fine-tuning of large language models for personality detection. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Lingzhi Shen, Yunfei Long, Xiaohao Cai, Guanming Chen, Yuhan Wang, Imran Razzak, and Shoaib Jameel. 2025c. Ll4g: Self-supervised dynamic optimization for graph-based personality detection. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Tommy Tandera, Derwin Suhartono, Rini Wongso, Yen Lina Prasetio, et al. 2017. Personality prediction system from facebook users. *Procedia computer science*, 116:604–611.

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

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Edward P Tighe, Jennifer C Ureta, Bernard Andrei L Pollo, Charibeth K Cheng, and Remedios de Dios Bulos. 2016. Personality trait classification of essays with the application of feature reduction. In *SAAIP@ IJCAI*, pages 22–28.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 91–103. Springer.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2023. Cdeval: A benchmark for measuring the cultural dimensions of large language models. *arXiv preprint arXiv:2311.16421*.

Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. 2023. Psycot: Psychological questionnaire as powerful chain-of-thought for personality detection. *arXiv preprint arXiv:2310.20256*.

Tao Yang, Feifan Yang, Haolan Ouyang, and Xiaojun Quan. 2021. Psycholinguistic tripartite graph network for personality detection. *arXiv preprint arXiv:2106.04963*.

Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. 2024. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023. Cultural compass: Predicting transfer learning success in offensive language detection with cultural features. *arXiv preprint arXiv:2310.06458*.

A Instruction Template Schema

Slot Category	Slot Type	Example Values
Cultural background	<culture>	Japanese, British, Arabic
	<culture_adj>	Japanese, British, Arabic (adjective form)
	<cultural_norm>	values politeness, encourages directness
	<language>	Japanese, English, Spanish
Personality traits	<trait_high>	high agreeableness, high conscientiousness
	<trait_low>	low neuroticism, low openness
	<trait_behavior>	calm and friendly, meticulous and careful
	<trait_emotion>	reserved and modest, enthusiastic and expressive
Role and task context	<role>	assistant, advisor, friend
	<task>	answer questions, provide suggestions
	<tone>	formal, casual, humorous, warm
	<formality>	formal expressions, conversational style

Table 5: Slot types and example values used in dynamic instruction templates.

B Prompt for Linguistic Evaluation

```
Based on the {conversation},
please evaluate the fluency and coherence of the answer
separately. You only need to judge whether the response is
natural and relevant. Each conversation follows the same
question-answer format.

Fluency: measures the readability and grammatical naturalness of
the response. The answer should be well-formed, clear, and easily
understandable to humans.

Coherence: evaluates the contextual and logical relevance of the
answer to the question. The answer should stay on topic, maintain
logical consistency, and avoid irrelevant or disjointed content.
```

Figure 4: Unified prompt used to evaluate the fluency and coherence of HiCAP's generated dialogues.

To ensure consistency and fairness across different LLM judges, we use a unified prompt to assess the fluency and coherence of HiCAP-generated dialogues. Figure 4 presents the complete prompt used in the evaluation.



Figure 5: Representative examples of HiCAP's culturally adaptive response generation under different personality configurations and interaction scenarios.

C Case Analysis

C.1 Culturally Adaptive Response Generation

Figure 5 presents six representative scenarios spanning Japanese, British, and Arabic personas with different personality configurations, social roles, and interaction contexts. The generated responses vary systematically in tone, formality, emotional expression, and communicative strategy according to these conditions.

For example, the Japanese personas with high agreeableness and low extraversion favour reassuring, modest, and socially restrained responses, as reflected in the supportive response to an anxious friend and the preference for small conversations at a birthday party. The British persona with high conscientiousness responds to workplace pressure with structured planning and a professional tone, while the high-openness traveller adopts a more exploratory style centred on trying local experiences and communicating flexibly. The Arabic personas with high agreeableness produce warm and empathetic responses in interpersonal situations, while the public-queue example combines politeness with assertive norm enforcement. These differences show that the same cultural setting can support distinct response styles as personality traits, social roles, and conversational contexts change.

These cases qualitatively illustrate how personality conditioning enables HiCAP to realise culturally adaptive responses across different social roles and conversational contexts.

C.2 Safety Risk Analysis

We evaluate the overall proportions of negative and dangerous responses produced by HiCAP on the CPED dataset. For each HiCAP variant, we sample generated replies under the CPED generation settings described in Section 4.1 and use the same three external LLM judges introduced

in Section 4.4.3, namely GPT-5, Claude-Sonnet-4.5, and Gemini-2.5-Pro, to determine whether a reply should be labelled as negative or dangerous. Negative responses refer to sentiment deterioration that may escalate social tension, while dangerous responses refer to content involving threats, severe moral violations, or safety-critical harm.

Model	Negative	Dangerous
HiCAP-Llama-3.1-8B	0.33	0.06
HiCAP-Qwen-2.5-14B	0.71	0.17
HiCAP-Phi-4-14B	0.49	0.12

Table 6: Proportions (%) of negative and dangerous responses across HiCAP variants on CPED.

As shown in Table 6, all HiCAP variants have low measured rates of negative and dangerous responses under this evaluation setting. HiCAP-Qwen-2.5 has slightly higher proportions, which may be related to its stronger reproduction of emotionally charged conversational styles and subtle sentiment shifts in Chinese drama contexts. HiCAP-Llama-3.1 shows the lowest rates, while HiCAP-Phi-4 lies between the two. More importantly, dangerous-response rates remain within a narrow and very low range of 0.06% to 0.17% across all evaluated variants. Together with the controllability results reported above, this suggests that HiCAP can support persona-conditioned stylistic variation while maintaining low measured safety risk under the evaluated setting. The differences across backbones further indicate that residual safety behaviour may also depend on the underlying model.