PCoKG: Personality-aware Commonsense Reasoning with Debate

Weijie Li[1], **Zhongqing Wang**[1*], **Guodong Zhou**[1]

[1]School of Computer Science and Technology, Soochow University
silverbeats@qq.com, wangzq@suda.edu.cn, gdzhou@suda.edu.cn

Abstract

Most commonsense reasoning models overlook the influence of personality traits, limiting their effectiveness in personalized systems such as dialogue generation. To address this limitation, we introduce the Personality-aware Commonsense Knowledge Graph (PCoKG), a structured dataset comprising 521,316 quadruples. We begin by employing three evaluators to score and filter events from the ATOMIC dataset, selecting those that are likely to elicit diverse reasoning patterns across different personality types. For knowledge graph construction, we leverage the role-playing capabilities of large language models (LLMs) to perform reasoning tasks. To enhance the quality of the generated knowledge, we incorporate a debate mechanism consisting of a proponent, an opponent, and a judge, which iteratively refines the outputs through feedback loops. We evaluate the dataset from multiple perspectives and conduct fine-tuning and ablation experiments using multiple LLM backbones to assess PCoKG's robustness and the effectiveness of its construction pipeline. Our LoRA-based fine-tuning results indicate a positive correlation between model performance and the parameter scale of the base models. Finally, we apply PCoKG to persona-based dialogue generation, where it demonstrates improved consistency between generated responses and reference outputs. This work bridges the gap between commonsense reasoning and individual cognitive differences, enabling the development of more personalized and context-aware AI systems.

Code — https://github.com/silverbeats/pcs_v2

1 Introduction

Commonsense reasoning remains a key challenge in machine intelligence (Storks, Gao, and Chai 2019). To advance this capability, NLP research has developed datasets such as ATOMIC (Sap et al. 2019; Hwang et al. 2021), which focus on if-then reasoning about events, including their causes (*xIntent*) and effects (*xEffect*). COMET (Bosselut et al. 2019), built on ATOMIC, has been applied to emotion recognition (Zhao, Zhao, and Lu 2022), empathetic dialogue generation (Wang et al. 2022; Sabour, Zheng, and Huang 2022; Tu et al. 2022), where contextual understanding is essential.

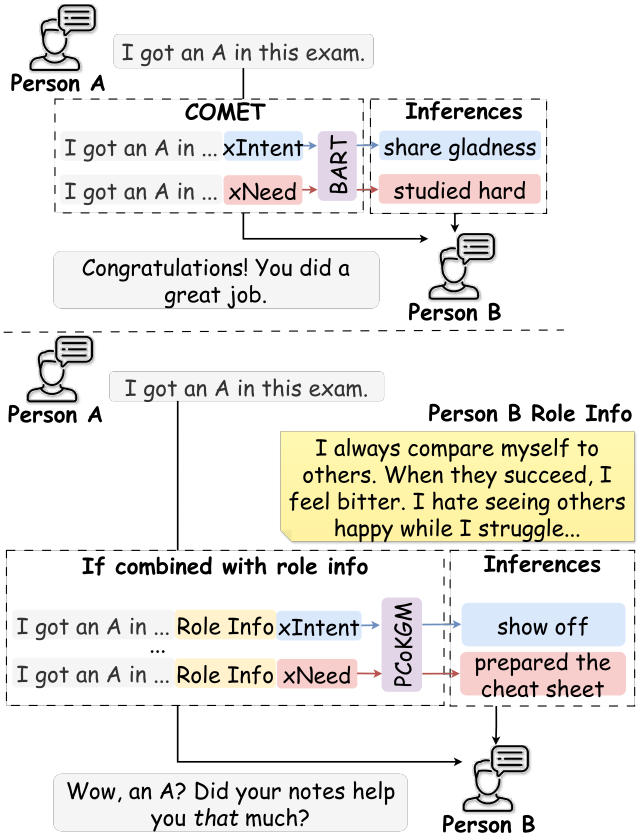

Figure 1: The upper panel shows how COMET is commonly used, while the lower panel illustrates a more realistic application incorporating personal traits.

Commonsense knowledge in ATOMIC is structured as (e, r, t) triplets, where e represents an event, r an inference dimension, and t the outcome. While useful for general reasoning, this framework overlooks individual differences—particularly personality traits—that shape how people interpret and respond to events. As a result, it fails to model real-world cognitive diversity and limits its applicability in personalized AI. As illustrated in the upper part of Figure 1, existing approaches like COMET generate plausible but generic inferences without considering personal

*Corresponding Author

traits. In contrast, the lower part illustrates the significant influence of individual perspectives, resulting in more diverse and personalized responses. Current frameworks, therefore, fall short in capturing the richness of human cognition.

To address this limitation and model individual differences in commonsense reasoning, we introduce the Personality-aware Commonsense Knowledge Graph (PCoKG), which enhances traditional knowledge graphs by integrating person-specific personality traits. We expand the conventional (e, r, t) triplet structure into a quadruple format: (e, p, r, t), where p represents personality information based on the Myers-Briggs Type Indicator (MBTI) (Myers 1987). This design enables more realistic modeling of how individuals with different personality types perceive and respond to the same event. For example, an introverted person may view social interactions very differently compared to an extroverted individual. By incorporating personality, our approach facilitates more personalized and context-aware inferences, enabling AI systems better to reflect the cognitive diversity inherent in human reasoning.

Constructing a high-quality commonsense knowledge graph enriched with personality information is challenging. Traditional crowdsourcing methods are limited by the difficulty and cost of recruiting individuals with specific personality traits at scale. To address this, we propose a novel LLM-driven approach that simulates diverse personality types using the MBTI framework. Our method introduces two key components: an evaluator-based mechanism that filters events and inference dimensions to ensure relevance and diversity, and a role-playing prompting strategy enhanced with a multi-agent debate mechanism to improve reasoning consistency and quality. Based on this approach, we construct PCoKG — the first large-scale, personality-aware commonsense knowledge graph, which contains $521,316$ high-quality (e, p, r, t) quadruples across 9 inference dimensions and all 16 MBTI types.

Compared to existing resources, PCoKG offers three major advantages. First, it is the first large-scale dataset and modeling framework that explicitly incorporates personality into commonsense reasoning, enabling personalized and contextually appropriate inferences. Second, our LLM-based pipeline eliminates the need for labour-intensive human annotation, supporting scalable data expansion. Third, although the dataset is built upon the MBTI framework, the pipeline is designed to be generalizable, allowing adaptation to other personality theories or the incorporation of additional role attributes. To validate the effectiveness of our approach, we conducted comprehensive experiments using multiple LLM backbones, demonstrating the robustness and generalisation of the generated knowledge across different architectures. We further performed extensive ablation studies to assess the contributions of each component in our pipeline and analyzed performance across model sizes to understand the impact of parameter scale. Finally, we applied PCoKG to a persona-based dialogue generation task, showcasing its practical utility in real-world applications. These results validate both the effectiveness of our methodology and the value of incorporating personality into commonsense knowledge modeling.

2 Related Work

2.1 Commonsense Knowledge Bases

ConceptNet (Speer, Chin, and Havasi 2017) is a widely used commonsense knowledge base (Commonsense Knowledge Base, CKB), though its Chinese version contains relatively limited knowledge (Kuo et al. 2009). To address such limitations, Zhang et al. (2021) automatically constructed a large-scale commonsense knowledge base called TransOMCS by transforming syntactic parses of web sentences into structured triples. However, most existing CKBs primarily focus on taxonomic relationships, such as *isA* and Synonym (Davis and Marcus 2015), which inevitably restricts their applicability. In contrast, another line of commonsense knowledge bases—represented by ATOMIC—requires human annotators to infer the causes and consequences of given events based on personal commonsense knowledge, thereby enriching the knowledge graph with tail entities (Sap et al. 2019; Hwang et al. 2021).

Inspired by ATOMIC's organisational structure, Li et al. (2022) introduced C^3KG, the first Chinese conversational commonsense knowledge graph built upon four types of conversational flows linking headers (events) and tails (inference results). Following the ATOMIC framework, Wang et al. (2024) built an emotional commonsense knowledge graph. Yang et al. (2024) constructed a Chinese dataset for personalized-aware commonsense reasoning, which is the most closely related to our work. However, their dataset was built manually, making it difficult to scale up in terms of size and coverage.

2.2 Myers-Briggs Type Indicator

The Myers-Briggs Type Indicator (MBTI) (Myers 1987) is a widely recognized personality model that categorizes individuals along four dichotomous dimensions: Introversion/Extraversion, Sensing/Intuition, Thinking/Feeling, and Judging/Perceiving. Despite psychometric critiques (Barbuto Jr 1997), it is widely used in both professional and personal contexts.

Due to its intuitive appeal, the MBTI has gained traction in computational research. Recent studies have leveraged MBTI for personality-aware natural language processing and human-computer interaction. Examples include MBTI-labeled Reddit datasets (Gjurković and Šnajder 2018), personalized emotional support systems (Tu et al. 2023), embedding stable MBTI traits into LLMs (Cui et al. 2023), simulating MBTI types to evaluate LLM decision-making, affective computing datasets integrating MBTI and emotion (Zhou, Luo, and Chen 2024), conversational MBTI and gender inference models (Shahnazari and Moein Ayyoubzadeh 2025), investigations into personality-adaptive dialogue agents (Cheng, Chang, and Chen 2025), and analyses of MBTI-induced bias in hate speech detection (Yuan et al. 2025). Collectively, these studies highlight the potential of MBTI-aware modeling to enhance personalisation, fairness, and behavioural fidelity in AI systems.

While existing commonsense knowledge bases primarily emphasise general causal and taxonomic relationships, and some recent efforts have included personality traits in

Algorithm 1 Event and Reasoning Dimension Acquisition

Initialize: $ER \leftarrow \emptyset$ {Set of event-dimension pairs}
E {List of events}, $EVLS$ {List of evaluators}, $CRIT$ {List of criteria}, $CRMap$ {Mapping from criterion to reasoning dimension}
for each $e \in E$ **do**
  for each $c \in CRIT$ **do**
    $r \leftarrow CRMap(c)$
    $S \leftarrow \emptyset$
    for each $evl \in EVLS$ **do**
      $s \leftarrow evl(e, c)$
      Append s to S
    end for
    if $\forall s \in S,\ s \geq 6$ **then**
      Append (e, r) to ER
    end if
  end for
end for
return ER

Algorithm 2 Personality-aware Debate Process

Input: Event e, reasoning dimension r, personality type p, inference model M
Output: Inference result t
Initialize empty histories: rp_hist, pro_hist, opp_hist, jdg_hist
for $i = 1$ to max_generate_times **do**
  $t \leftarrow M(e, r, p)$
  for $round = 1$ to debate_rounds **do**
    $pro\_resp \leftarrow \textsc{Pro}(e, r, p, t, pro\_hist)$
    Append pro_resp to pro_hist and opp_hist
    $opp\_resp \leftarrow \textsc{Opp}(e, r, p, t, opp\_hist)$
    Append opp_resp to pro_hist and opp_hist
    Append both responses to jdg_hist
  end for
  $jdg\_resp \leftarrow \textsc{Jdg}(e, r, p, jdg\_hist)$
  if jdg_resp is acceptable **then**
    return t
  else
    Append jdg_resp to rp_hist for feedback
  end if
end for
return None

NLP tasks, our work distinguishes itself by systematically incorporating personality-aware reasoning into a large-scale commonsense knowledge graph. Unlike manually curated datasets, which often have limitations in scale and coverage, we propose a fully scalable framework that leverages LLMs to simulate diverse MBTI personality types in commonsense reasoning tasks.

3 Construction of the Personality-aware Commonsense Knowledge Graph

The data construction pipeline consists of two stages, and its key steps are detailed in Figure 2.

First, we enumerate event–reasoning pairs (e, r) from the ATOMIC knowledge base and filter them using three LLM evaluators, which score each pair on a 10-point scale for its potential to elicit personality-diverse responses. Only pairs scoring above 6 from all evaluators are retained.

Second, we sample MBTI personality types p according to their global population distribution[1] and prompt an LLM to role-play each selected (e, r) pair, generating personalized reasoning outputs t. Each instance in PCoKG is thus a quadruple (e, r, p, t) (see examples in Table 1).

The following sections detail these two stages: (1) acquisition of events and reasoning dimensions, and (2) personality-conditioned reasoning generation.

3.1 Event and Reasoning Dimension Acquisition

Initially, events are extracted from ATOMIC$^{20}_{20}$ (Hwang et al. 2021). To ensure linguistic quality, we employ *language_tool_python* to filter out events containing grammatical errors, resulting in a final set of $19,184$ well-formed events. Subsequently, we select three large language models—Deepseek-R1, Qwen-Turbo, and Doubao-1.6-Seed—as evaluation models. We define nine evaluation criteria centred on the reasoning dimension, aiming to assess whether each event applies to reasoning processes associated with different personality types. Each criterion is rated on a 10-point scale. The overall extraction algorithm is presented in Algorithm 1. The specific evaluation criteria are as follows. Finally, we obtain $95,783$ pairs of (e, r), involving $15,227$ events, denoted as $\mathcal{E}$.

- **xIntent (Motivation)** Does the event lead to clearly different internal drives depending on the character's MBTI type?
- **xWant (Plan)** Do different MBTI types form different plans or intentions in response to the event?
- **xEffect (Impact)** Does the event have different psychological or behavioural impacts on different MBTI types?
- **xReact (Emotional Response)** Do different MBTI types show different emotional reactions to the event?
- **xNeed (Preparation)** Would different MBTI types prepare differently for this event?
- **xAttr (Self-Narration)** How would different MBTI types describe this event in first person? Is there variation in tone or perspective?
- **oReact (Others' Emotion)** Do different MBTI types make different assumptions about how others feel about the event?
- **oWant (Others' Intention)** Do different MBTI types expect others to react differently to the event?
- **oEffect (Impact on Others)** Do different MBTI types assume different levels of impact on others due to the event?

3.2 Personality-aware Reasoning via Debate

We guide large language models to adopt roles corresponding to different MBTI personality types and perform rea-

[1] https://www.16personalities.com/country-profiles/global/world#global

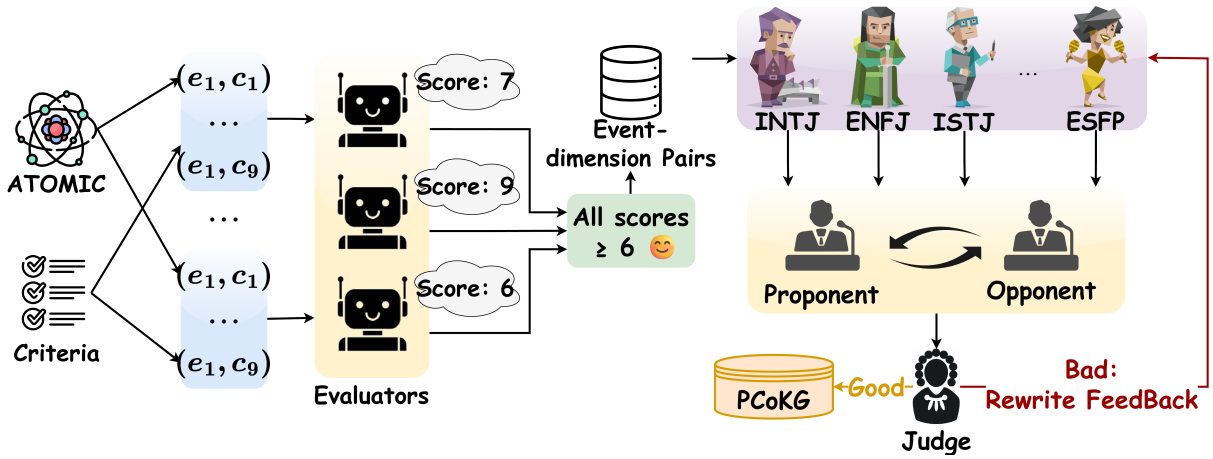

Figure 2: PCoKG construction process.

Event	Dimension	MBTI	Inference
PersonX makes any money	xWant	ISFJ	Plan to save some and use the rest for family and friends
		INTP	Dive deeper into the theoretical aspects to uncover underlying principles
PersonX locks PersonY's keys in PersonY's car	xIntent	ENFP	Just for fun and to see PersonY's reaction
		INTJ	To assess PersonY's response time

Table 1: Samples of PCoKG data.

soning based on the entity-reasoning set ($\mathcal{E}$) obtained in the previous section. To enhance model comprehension and the quality of inferences, we convert each reasoning dimension into clear, human-readable natural language descriptions.

To further improve the reliability of the generated inferences, we introduce a debate framework, with the core algorithm detailed in Algorithm 2. In this framework, we define three roles: **Proponent**, **Opponent**, and **Judge**. The Proponent argues that the model's reasoning aligns with the target MBTI type and provides supporting evidence. In contrast, the Opponent challenges the consistency between the reasoning and the expected type.

Multiple rounds of debate are held between the Proponent and Opponent, after which the Judge evaluates their arguments and delivers a final judgment. If the initial reasoning does not meet the desired standard, the Judge offers feedback and suggestions for improvement, prompting the model to refine its output iteratively.

4 Dataset Analysis

The basic statistical information of PCoKG is shown in Table 2. To assess whether the LLM-generated reasoning responses are meaningfully aligned with the assigned MBTI types, we evaluate the dataset through three lenses: Readability-Personality Association, Personality-Reasoning Association, and Human Evaluation.

4.1 Readability-Personality Association Analysis

Personality traits are known to influence language use and communication styles. This study uses the Flesch Reading Ease score—a standard measure of text readability (ranging from 0 to 100, with higher scores indicating simpler language)—to evaluate linguistic complexity across MBTI types.

Metric	Value
Data Size	521,316
Number of Events	15,077
Average Event Length	4.79
Average Reasoning Outcome	8.75

Table 2: Statistical summary of PCoKG.

Figure 3 shows notable differences in readability across MBTI types. For example, ESFP (77.7) and ESTP (74.0) have high readability scores, suggesting a preference for direct, concrete expression. In contrast, INTJ (37.0) and INTP (39.6) score much lower, reflecting more complex and abstract language. These findings align with MBTI theory: Thinking (T) and Intuitive (N) types favour logical and abstract reasoning, while Feeling (F) and Sensing (S) types prefer emotional and accessible language.

Overall, the results suggest that LLM-generated reasoning content reflects linguistic traits consistent with the target MBTI type, supporting its effectiveness in personality perception tasks.

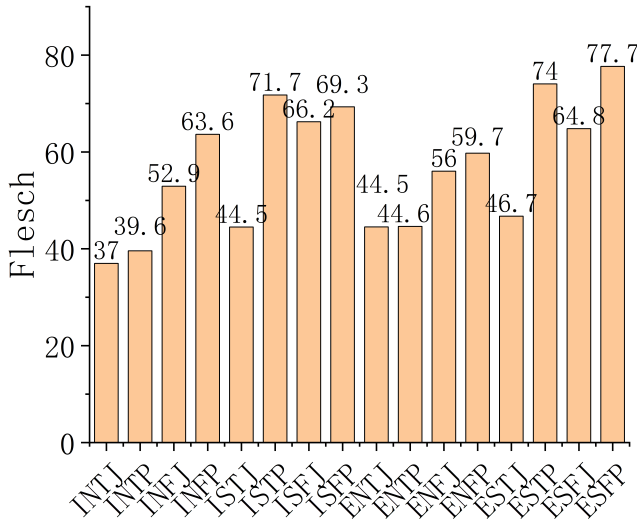

Figure 3: Flesch reading ease scores across 16 MBTI personality types.

Dimension	AMI	AMI Shuffle
xReact	0.255950	-0.000009
xWant	0.237505	0.000021
xIntent	0.240094	-0.000033
xNeed	0.177847	-0.000029
xEffect	0.216289	-0.000026
xAttr	0.511176	-0.000027
oWant	0.121604	0.000062
oEffect	0.161030	-0.000008
oReact	0.114518	-0.000044

Table 3: Adjusted MI (AMI) between MBTI personality types and clustering labels of reasoning results (p-value $<$ 0.01).

4.2 Personality-Reasoning Association Analysis

To evaluate whether LLM-generated reasoning reflects meaningful MBTI-aligned signals, we perform mutual information analysis. Reasoning styles are derived via K-means clustering on sentence embeddings, with cluster numbers selected based on Silhouette Scores. We compute Adjusted Mutual Information (AMI) between cluster IDs and MBTI types, which accounts for chance agreement and offers a reliable comparison. As a baseline, we shuffle MBTI labels 100 times and recalculate AMI, yielding near-zero null values (Ojala, Pietikainen, and Maenpaa 2002). Mann-Whitney U tests show all original AMI values are significantly higher ($p < 0.01$), confirming strong associations. Results (Table 3) indicate that dimensions related to self-perception (e.g., *xAttr*, *xIntent*) align more clearly with MBTI types, suggesting that LLM-generated reasoning varies meaningfully across personality types.

4.3 Human Evaluation

To further validate the presence of personality-aware reasoning signals in our dataset, we conducted a human evaluation

Metric	Average Score
Personality Consistency	1.63
Reasoning Coherence	1.78
Naturalness	1.71

Table 4: Human evaluation results.

in which three psychology graduate students familiar with MBTI assessed $1,440$ randomly selected instances covering all 16 MBTI types and 9 reasoning dimensions. Each reasoning result received ratings on three criteria: personality consistency, reasoning coherence, and naturalness, using a 3-point scale where 0 indicates *No*, 1 indicates *Somewhat*, and 2 indicates *Yes*. The average inter-annotator agreement, measured by Fleiss' Kappa, was 0.57, indicating moderate to substantial agreement. The results summarized in Table 4 indicated that the reasoning responses were generally coherent, with an average score of 1.78, and natural, with an average score of 1.71. However, the average score for personality consistency was slightly lower at 1.63. This suggested that while most responses aligned with the intended personality type, there was still room for improvement.

5 Experiments

5.1 Setup

Dataset To evaluate the effectiveness of the PCoKG dataset in supporting personalized commonsense reasoning, we partition the data into training, validation, and test sets using a 9:0.5:0.5 split based on events. This resulted in $13,576$ training events ($468,479$ quadruples), 750 validation events ($26,321$ quadruples), and 751 test events ($26,516$ quadruples).

Metrics We adopt word-overlap metrics—specifically, BLEU-4 (**B-4**), Rouge-1 (**R-1**), Rouge-2 (**R-2**), and Rouge-L (**R-L**)—to assess the degree of lexical alignment between the generated outputs and the reference answers. These metrics provide a quantitative measure of how well the model reproduces content that resembles the expected responses in terms of n-gram overlap.

Models We select three base models, Qwen3-0.6B, LLaMA3-1B, and MiniCPM4-0.5B, and conduct full-parameter fine-tuning on the PCoKG dataset to equip them with personalized commonsense reasoning capabilities. We refer to the resulting fine-tuned models collectively as **PCoKGM**. For comparison, we also train the **COMET** model, into which the reasoning dimensions and personality types are incorporated as special tokens in the tokenizer. The input to the model is constructed by concatenating three components: the event, the reasoning dimension, and the personality type, to predict the corresponding reasoning outcome. Additionally, we design one-shot prompting templates to evaluate the performance of Deepseek-R1 (**R1**), Doubao-seed-1.6-thinking (**1.6-Thinking**), and GPT-o4-mini (**o4-mini**) on the PCoKG task.

Model	B-4	R-1	R-2	R-L
R1	2.67	14.45	1.89	13.44
1.6-Thinking	3.39	15.43	2.45	13.96
o4-mini	5.38	15.34	2.09	14.28
COMET				
- LLaMA3	12.58	30.51	12.77	28.91
- Qwen3	10.09	26.31	9.26	25.00
- MiniCPM4	10.81	28.29	10.43	26.78
PCoKGM				
- LLaMA3	13.73	32.09	14.31	30.53
- Qwen3	14.08	32.68	14.78	31.07
- MiniCPM4	**14.50**	**32.99**	**15.27**	**31.38**

Table 5: Comparison of model performance on PCoKG. Best results are indicated in **bold**.

Implementation Details We fine-tune the models using `LLaMA-Factory` on four 3090 GPUs. Each GPU is assigned a batch size of 8, with gradient accumulation over 4 steps. The warmup ratio is set to 0.1, and the cosine learning rate scheduler is employed. The models are trained for one epoch on the training set, with validation performance evaluated every 300 training steps. Early stopping is applied if the performance on the validation set does not improve for three consecutive evaluations.

5.2 Comparative Performance Analysis

As shown in Table 5, PCoKGM significantly outperforms all baseline models across all metrics, achieving the highest scores. This demonstrates that incorporating reasoning dimensions and personality types as natural language prompts into the input enhances the model's ability to generate contextually relevant and personalized commonsense knowledge. In contrast, the COMET model encodes these signals as special tokens rather than interpretable natural language. While this approach enables structured control over reasoning patterns, it may limit interpretability and generalization in downstream applications.

Among the evaluated large language models, the three models exhibit comparable performance, yet all fall short of PCoKGM. This highlights the advantages of domain-specific fine-tuning and the explicit integration of personalization signals into the model's input structure. Moreover, the generated outputs from these LLMs indicate that, although they are capable of role-playing to some extent, their initial generations do not fully align with the desired outcomes. This further validates the necessity of employing a debate framework during PCoKG construction, where generations are refined through feedback from a judge model.

5.3 Ablation Experiment Analysis

To further validate the effectiveness of each component in our dataset construction, we design four ablation settings: *w/o mbti*, *w/o select*, *w/o debate*, and *w/o select & debate*. The results, summarized in Table 6, highlight the distinct contribution of each module.

Base Model	B-4	R-1	R-2	R-L
LLaMA3	**13.73**	**32.09**	**14.31**	**30.53**
- w/o mbti	10.16	25.59	9.36	24.51
- w/o select	11.25	27.92	10.59	26.49
- w/o debate	12.09	29.45	12.04	28.08
- w/o select & debate	10.66	26.00	9.62	24.72
Qwen3	**14.08**	**32.68**	**14.78**	**31.07**
- w/o mbti	10.33	25.72	9.55	24.68
- w/o select	10.56	28.72	9.38	24.53
- w/o debate	11.00	26.84	9.96	25.62
- w/o select & debate	10.66	26.60	9.78	25.28
MiniCPM4	**14.50**	**32.99**	**15.27**	**31.38**
- w/o mbti	10.45	25.94	9.66	24.88
- w/o select	10.89	27.33	10.21	25.91
- w/o debate	11.46	27.46	10.68	26.25
- w/o select & debate	10.93	26.51	10.14	25.19

Table 6: The ablation experimental results of PCoKGM using different base models. Best results are in **bold**.

Removing MBTI personality types (*w/o mbti*) leads to the largest performance drop across all models and metrics, underscoring their critical role in structured reasoning. Removing the selection mechanism (*w/o select*) also reduces performance, indicating that while learning from raw data is possible, curated high-quality event–dimension pairs enhance reasoning accuracy. The *w/o debate* setting results in a moderate decline, suggesting that the debate framework enhances reasoning depth and ensures consistency in personality-aligned responses. The worst performance occurs when both selection and debate are removed (*w/o select & debate*), confirming their complementary roles.

In summary, each component plays a distinct role: MBTI provides the reasoning anchor, selection ensures data quality, and debate enhances the consistency of personality-aware reasoning. These findings highlight the value of a structured, multi-stage dataset construction process for personality-driven reasoning tasks.

5.4 Impact of Model Scale

We evaluate three model families—Qwen3, LLaMA3, and MiniCPM4—across multiple scales to study how model size affects performance on our personality-aware commonsense reasoning dataset. As shown in Figure 4, performance improves with scale across all models, suggesting that larger models better capture the nuanced relationships between personality types, events, and reasoning outcomes. This trend is consistent across BLEU-4 and ROUGE metrics, indicating enhanced fluency and coherence in generated inferences as model capacity increases. Notably, the performance gap between model families diminishes at larger scales, implying that sufficient model size may compensate for architectural or training differences. Additionally, improvements in ROUGE-1, ROUGE-2, and ROUGE-L largely align with the BLEU-4 trends, reinforcing the notion that larger models are better equipped to handle complex, personality-conditioned reasoning tasks. These findings underscore the significance of model scale in discerning fine-grained behav-

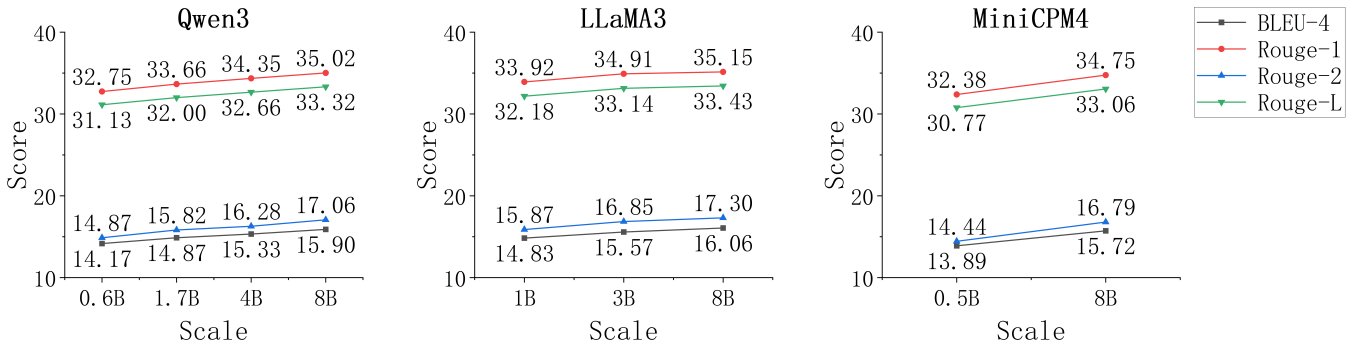

Figure 4: Performance of the foundation model with varying sizes on the PCoKG under LoRA fine-tuning.

ioral patterns from large-scale structured data. Overall, the results suggest that scaling up model size is a viable strategy for improving reasoning accuracy and coherence, especially when dealing with multi-dimensional, personality-sensitive inference tasks.

5.5 Application

To validate the effectiveness of our proposed PCoKG in practical applications, we conduct experiments on a dialogue generation task. Specifically, we utilize the SPC dataset's test set (967 samples) (Jandaghi et al. 2024) to evaluate how well our approach enhances personalized reasoning compared to existing methods.

Initially, we trained a BERT classifier on the *MBTI Personality Types 500 Dataset*[2] from Kaggle. This classifier was then employed to predict the MBTI personality types of the conversational participants in the SPC dataset based on their character descriptions. Using these predictions, PCoKGM—fully fine-tuned on MiniCPM4-0.5B—generates responses that integrate context and personality. For comparison, we also evaluate COMET, which relies solely on contextual content for inference.

The results, summarized in Table 7, reveal several key insights. First, integrating commonsense reasoning into dialogue generation yields better outcomes. Both COMET and PCoKGM outperform models without commonsense reasoning across various metrics, indicating that leveraging commonsense knowledge significantly enhances the alignment between generated outputs and expected responses. Second, incorporating personality traits into commonsense reasoning further improves the quality of generated dialogues. Across all tested models, PCoKGM consistently outperforms COMET. These findings suggest that modeling personality-specific nuances leads to more accurate and engaging dialogues.

6 Conclusion

This study introduces Personality-aware Commonsense Knowledge Graph (PCoKG), a knowledge graph enhanced with personality traits to support personalized, context-aware reasoning. It integrates event quality control, reasoning dimensions, and a role-based debate mechanism to build a large-scale knowledge graph. Experiments confirm the effectiveness of each component and explore the impact of model scale. We also demonstrate PCoKG's value in dialogue generation. Limitations include a focus on personality traits alone, without considering factors like gender or occupation. Future work will incorporate these aspects to develop a more comprehensive framework for personalized reasoning.

[2]https://www.kaggle.com/datasets/zeyadkhalid/mbti-personality-types-500-dataset

Model	B-4	R-1	R-2	R-L
R1	10.40	26.04	10.88	24.88
- w/ COMET	11.76	28.58	13.24	27.40
- w/ PCoKGM	**12.03**	**29.69**	**13.89**	**28.47**
1.6-Thinking	16.51	37.41	20.17	35.80
- w/ COMET	18.28	40.31	22.19	38.92
- w/ PCoKGM	**18.92**	**40.89**	**22.85**	**39.50**
o4-mini	6.05	18.55	5.69	17.63
- w/ COMET	6.65	20.24	6.84	19.16
- w/ PCoKGM	**6.85**	**20.27**	**7.09**	**19.31**

Table 7: Comparison of dialogue generation performance. Best results are in **Bold**.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62376178), Jiangsu Key Laboratory of Language Computing(JSLCKeyLab 202500003), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Barbuto Jr, J. E. 1997. A critique of the Myers-Briggs Type Indicator and its operationalization of Carl Jung's psychological types. *Psychological Reports*, 80(2): 611–625.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proc. ACL*, 4762–4779. Florence, Italy: ACL.

Cheng, S.; Chang, W.-Y.; and Chen, Y.-N. 2025. Exploring Personality-Aware Interactions in Salesperson Dialogue Agents. arXiv:2504.18058.

Cui, J.; Lv, L.; Wen, J.; Wang, R.; Tang, J.; Tian, Y.; and Yuan, L. 2023. Machine mindset: An mbti exploration of large language models. arXiv:2312.12999.

Davis, E.; and Marcus, G. 2015. Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence. *Commun. ACM*, 58(9): 92–103.

Gjurković, M.; and Šnajder, J. 2018. Reddit: A Gold Mine for Personality Prediction. In *Proc. Workshop on PEOPLES*, 87–97. New Orleans, Louisiana, USA: ACL.

Hwang, J. D.; Bhagavatula, C.; Le Bras, R.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2021. (Comet-) atomic 2020: on symbolic and neural commonsense knowledge graphs. In *Proc. AAAI*, 6384–6392. Virtual Event: AAAI Press.

Jandaghi, P.; Sheng, X.; Bai, X.; Pujara, J.; and Sidahmed, H. 2024. Faithful Persona-based Conversational Dataset Generation with Large Language Models. In *Proc. Workshop on NLP4ConvAI*, 114–139. Bangkok, Thailand: ACL.

Kuo, Y.-l.; Lee, J.-C.; Chiang, K.-y.; Wang, R.; Shen, E.; Chan, C.-w.; and Hsu, J. Y.-j. 2009. Community-Based Game Design: Experiments on Social Games for Commonsense Data Collection. In *Proc. ACM SIGKDD Workshop on Human Computation*, 15–22. Paris, France: ACM.

Li, D.; Li, Y.; Zhang, J.; Li, K.; Wei, C.; Cui, J.; and Wang, B. 2022. C^3KG: A Chinese Commonsense Conversation Knowledge Graph. In *Proc. Findings of ACL*, 1369–1383. Dublin, Ireland: ACL.

Myers, I. B. 1987. *Introduction to type: A description of the theory and applications of the Myers-Briggs Type Indicator*. Consulting Psychologists Press.

Ojala, T.; Pietikainen, M.; and Maenpaa, T. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 24(7): 971–987.

Sabour, S.; Zheng, C.; and Huang, M. 2022. Cem: Commonsense-aware empathetic response generation. In *Proc. AAAI*, 11229–11237. Virtual Event: AAAI Press.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proc. AAAI*, 3027–3035. AAAI Press.

Shahnazari, K.; and Moein Ayyoubzadeh, S. 2025. Who Are You Behind the Screen? Implicit MBTI and Gender Detection Using Artificial Intelligence. arXiv:2503.09853.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proc. AAAI*, 4444–4451. San Francisco, California, USA: AAAI Press.

Storks, S.; Gao, Q.; and Chai, J. Y. 2019. Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. arXiv:1904.01172.

Tu, Q.; Chen, C.; Li, J.; Li, Y.; Shang, S.; Zhao, D.; Wang, R.; and Yan, R. 2023. Characterchat: Learning towards conversational ai with personalized social support. arXiv:2308.10278.

Tu, Q.; Li, Y.; Cui, J.; Wang, B.; Wen, J.-R.; and Yan, R. 2022. MISC: A Mixed Strategy-Aware Model integrating COMET for Emotional Support Conversation. In *Proc. ACL*, 308–319. Dublin, Ireland: ACL.

Wang, L.; Li, J.; Lin, Z.; Meng, F.; Yang, C.; Wang, W.; and Zhou, J. 2022. Empathetic Dialogue Generation via Sensitive Emotion Recognition and Sensible Knowledge Selection. In *Proc. Findings of EMNLP*, 4634–4645. Abu Dhabi, United Arab Emirates: ACL.

Wang, Z.; Liu, X.; Hu, M.; Ying, R.; Jiang, M.; Wu, J.; Xie, Y.; Gao, H.; and Cheng, R. 2024. ECoK: Emotional Commonsense Knowledge Graph for Mining Emotional Gold. In *Proc. Findings of ACL*, 8055–8074. Bangkok, Thailand: ACL.

Yang, Y.; Li, W.; Fan, X.; Deng, W.; Liu, J.; Diao, Y.; and Tuerxun, P. 2024. Chinese Personalized Commonsense Understanding and Reasoning Based on Curriculum-Learning. In *Proc. NLPCC*, 213–225. Hangzhou, China: Springer Nature Singapore.

Yuan, S.; Nie, E.; Tawfelis, M.; Schmid, H.; Schütze, H.; and Färber, M. 2025. Hateful Person or Hateful Model? Investigating the Role of Personas in Hate Speech Detection by Large Language Models. arXiv:2506.08593.

Zhang, H.; Khashabi, D.; Song, Y.; and Roth, D. 2021. TransOMCS: from linguistic graphs to commonsense knowledge. In *Proc. IJCAI*, 4004–4010. Virtual Event: IJCAI Organization.

Zhao, W.; Zhao, Y.; and Lu, X. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *Proc. IJCAI*, 4524–4530. Vienna, Austria: IJCAI Organization.

Zhou, J.; Luo, S.; and Chen, H. 2024. A Chinese Multi-label Affective Computing Dataset Based on Social Media Network Users. arXiv:2411.08347.