ProtoMBTI: Prototype-Guided Retrieval-Augmented Reasoning for MBTI Inference from Text

Anonymous ACL submission

Abstract

Understanding a user's personality is crucial for personalized AI, and the MBTI provides a widely recognized operational framework for personality modeling. Existing text-based MBTI prediction methods often treat labels as fixed categories, neglecting the prototype-based nature of personality emphasized in cognitive psychology. To address this, we propose **ProtoMBTI**[1], a prototype-based reasoning framework for social-media text: it aligns LLM inference with the cognitive structure of personality via prototype retrieval-driven reasoning. Specifically, PROTOMBTI constructs a balanced, high-quality prototype library and performs a retrieve–reuse–revise–retain cycle during inference to achieve accurate, interpretable, and transferable predictions. On the Kaggle (85.14%) and Pandora (71.41%) benchmarks, PROTOMBTI significantly outperforms neural and LLM baselines, and under distribution shift achieves an average accuracy of 96.41% on the Pandora test set, covering all 16 MBTI types.

1 Introduction

Understanding a user's personality is an important enabler of personalized AI: it allows systems to adapt content and interaction style to individuals rather than rely on one-size-fits-all heuristics. In education, personality-aware tutors can adjust pacing and feedback framing to sustain engagement (Sajja et al., 2023); in recommendation, personality signals inferred from everyday text can help mitigate cold-start and disambiguate intent when behavioral history is sparse (He et al., 2018; Li et al., 2023); and in organizational settings, personality-sensitive analysis of internal communication can support team formation while accounting for individual styles (Wang, 2024). Across these settings, a common requirement is to infer relatively stable, person-level dispositions directly from language produced in the wild, without administering standalone psychometric tests.

In this work we focus on inferring personality using the Myers-Briggs Type Indicator (MBTI), a widely adopted typology in practice and online communities, and a common operational target in NLP benchmarks (Myers, 1962; Gjurković et al., 2020). Given a user's text, the task is to predict the user's MBTI type (four dichotomies yielding sixteen types) by extracting latent trait signals from user-generated language (Pan and Zeng, 2023; Li et al., 2025). Prior approaches span lexicon-based feature engineering, end-to-end neural models, and LLM-based methods using prompting or few-shot inference (Taboada et al., 2011; Ryan et al., 2023; Ashraf et al., 2024; Li et al., 2024; Hu et al., 2024; Li et al., 2025). While MBTI's psychometric status is debated in psychology, it remains a practical and widely used code for studying personality signals in text (McCrae and Costa, 1989).

Two limitations of most existing MBTI predictors are that (i) they treat personality types as fixed categorical targets, optimizing direct label prediction with little explicit structure for reasoning from representative evidence(Bi et al., 2025; Li et al., 2025, 2024; Patil et al., 2024; Zhu et al., 2024b; Shobha et al., 2024); and (ii) MBTI-related text in social media exhibits substantial noise, class imbalance, and cross-style variability, with extensive overlap in language use across personality types, making the assumption of strictly discrete labels difficult to sustain in practice(Hu et al., 2024). Motivated by prototype-based views of categorization(figure 1), where judgments are made by comparison to central exemplars rather than strict rules (Rosch and Mervis, 1975), we ask whether prototype-guided inference can provide a stronger and more interpretable basis for MBTI prediction from text. Recent work suggests that conditioning

[1]Our code is released anonymously at https://anonymous.4open.science/r/ProtoMBTI-6F00.

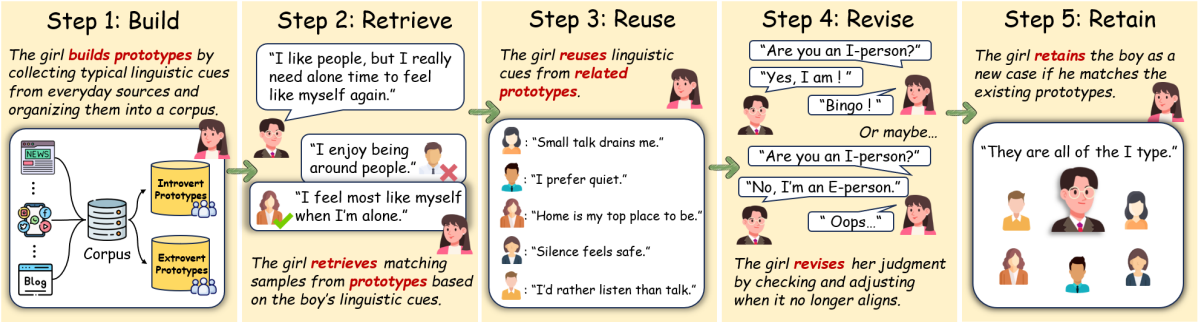

Figure 1: The figure shows how people form, apply, revise, and accumulate category judgments in daily life by recalling similar past cases and adjusting their decisions as new information becomes available, reflecting a case-based and prototype-driven cognitive process rather than a specific computational model.

LLMs on prototypical examples can improve task performance and make predictions more evidence-grounded (Zhu et al., 2024a; Deng et al., 2024; Wei et al., 2025; He et al., 2025; Ren et al., 2024). We therefore investigate: (1) how to construct an operational set of MBTI prototypes from text, and (2) how to retrieve and integrate these prototypes during inference to obtain predictions that are accurate, evidence-grounded, and robust across datasets. We study these questions in MBTI inference from social-media posts.

To address these questions, we propose ProtoMBTI, a prototype-guided framework for MBTI inference from social-media text. ProtoMBTI replaces direct label prediction with retrieve-then-reason inference: it retrieves representative prototypes from a curated prototype bank using a learned embedding space, and conditions an LLM on these prototypes to generate evidence-grounded dichotomy (and type) predictions. We construct the prototype bank with quality-controlled augmentation to reduce imbalance, and filter noisy synthetic samples using a four-dimension agreement criterion. Our main contributions are: (i) a quality-controlled prototype bank construction pipeline for MBTI text, (ii) a prototype-guided retrieval-and-reasoning inference procedure, and (iii) empirical and qualitative analyses demonstrating accuracy gains and evidence grounding. Empirically, we evaluate on two standard MBTI text benchmarks, Kaggle (Pan and Zeng, 2023) and Pandora (Gjurković et al., 2020), and report both average dichotomy accuracy and 16-type accuracy under a consistent protocol. ProtoMBTI achieves 85.14% (Kaggle) and 71.41% (Pandora) average dichotomy accuracy, improving over the strongest prior baseline (ETM) by +7.35 and +5.64 points; in a cross-dataset setting, it attains 90.87% average dichotomy accuracy and 81.15% 16-type accuracy on Pandora, suggesting improved robustness under domain mismatch.

2 Related Work and Main Motivation

2.1 Predicting MBTI Types from Text

Automatically detecting personality from text has become an increasingly prominent topic in computational psycholinguistics. Although the Big Five framework dominates psychological research (John and Srivastava, 1999), the Myers-Briggs Type Indicator (MBTI) (Myers, 1962; McCrae and Costa, 1989) remains widely used in online communities, self-assessment platforms, and workplace settings (Quenk, 1999). MBTI categorizes individuals into 16 types based on four dichotomies: **Extraversion vs. Introversion** (E/I; outward- vs. inward-oriented engagement), **Sensing vs. Intuition** (S/N; preference for concrete details vs. abstract patterns), **Thinking vs. Feeling** (T/F; analytic vs. value-oriented decision making), and **Judging vs. Perceiving** (J/P; structured planning vs. flexible adaptation).

Existing MBTI prediction approaches can be broadly grouped into three families. *Lexicon-based methods* extract handcrafted psycholinguistic features (e.g., LIWC-style dictionaries) and use conventional classifiers for prediction (Taboada et al., 2011; Komisin and Guinn, 2011; Tadesse et al., 2018). *Deep learning methods* employ CNNs, RNNs, and Transformers to learn representations from raw text end-to-end, typically improving over lexicon-based approaches (Xue et al., 2018; Tandera et al., 2017; Ryan et al., 2023; Ashraf et al., 2024; Shanmukha et al., 2024), and can be further enhanced using hierarchical modeling or auxiliary signals (Jiang et al., 2020; Keh et al.,

2019; Patil et al., 2024; Zhang, 2023; Tareaf, 2022; Shobha et al., 2024; Lynn et al., 2020; Zhu et al., 2024b; Yang et al., 2022; Ma et al., 2022; Yang et al., 2021a). More recently, *large language model (LLM)-based methods* achieve zero-shot or low-resource prediction via prompting, few-shot learning, or data augmentation (Li et al., 2024; Hu et al., 2024; Li et al., 2025; Bi et al., 2025). Despite steady progress, most approaches formulate MBTI inference as *direct label prediction*, with limited explicit use of representative evidence at inference time, which can make decisions harder to interpret and potentially brittle across data sources.

2.2 Prototype Theory as Design Motivation

Prototype theory (Rosch, 1973; Rosch and Mervis, 1975; Rosch, 1975) describes categorization as comparison to central, highly representative exemplars, with other instances judged by graded similarity rather than strict rules. We use this view as design motivation for MBTI inference from text: for example, within the E/I dichotomy, prototypical expressions such as "I enjoy going out with friends" or "I prefer solitude and reflection" can be treated as representative textual cues for the underlying category(see more in Appendix B). This perspective suggests an operational alternative to direct label prediction: (i) construct a library of representative prototypes that capture typical traits of each MBTI category, and (ii) infer labels for a new input by retrieving similar prototypes and reasoning from their cues. Recent studies also indicate that conditioning LLMs on prototypical examples can improve task performance and make outputs more evidence-grounded (Zhu et al., 2024a; Deng et al., 2024; Wei et al., 2025; He et al., 2025; Ren et al., 2024).

2.3 Closest Technical Relatives: Exemplar Retrieval and Case-Based Reasoning

Our retrieve-and-reason formulation is related to exemplar-based methods in NLP, including retrieval-augmented prompting and example selection for in-context learning, which use retrieved instances to guide prediction. It is also reminiscent of *case-based reasoning* (CBR), which solves new problems by retrieving similar past cases and adapting their solutions (Aamodt and Plaza, 1994; Kolodner, 1992; Wiratunga et al., 2024; Hatalis et al., 2025). In contrast to classical CBR pipelines, our focus is on an LLM-centered instantiation where retrieved prototypes serve as *evidence* for personality inference, and where the prototype library is explicitly constructed and filtered for quality in the MBTI setting.

3 Task Formulation

We view MBTI types as a structured label space composed of four binary axes. Let $\mathcal{C}^{(i)} = \{0, 1\}$ for $i \in \{1, 2, 3, 4\}$ denote the label set for the i-th MBTI *dimension*, where the two labels correspond to: E/I, S/N, T/F, J/P. An MBTI *type* is the 4-tuple of choices across these axes, yielding the 16-type space $\mathcal{C}^{\mathrm{MBTI}} = \mathcal{C}^{(1)} \times \mathcal{C}^{(2)} \times \mathcal{C}^{(3)} \times \mathcal{C}^{(4)}$.

A prototype p is defined as a structured representation $p = \langle a, e, c \rangle$, where a denotes the *attribute* (e.g., a user-generated text), $e \in \mathbb{R}^d$ denotes the *embedding* encoding the relational features between a and the category c, and $c \in \mathcal{C}^{\mathrm{MBTI}}$ denotes a MBTI personality type. Although the label space has a natural dimension-level factorization, we treat 16-type prediction as the primary task; the dimension labels are used only to guide augmentation and to validate prototypes during quality control.

Task Definition. Given a user-generated text a, the objective is to infer the author's MBTI personality type $c \in \mathcal{C}^{\mathrm{MBTI}}$.

4 Methodology

ProtoMBTI proposes a two-stage prototype-driven framework for personality inference, where prototype quality is central to effective reasoning. However, existing personality datasets are insufficient for constructing strong and representative prototypes due to limited scale, severe class imbalance, and weak expressive diversity (Appendix G). We therefore enhance the data to build a class-balanced and expressive prototype library. Based on these prototypes, ProtoMBTI adopts a CBR-like paradigm to align test instances with historical prototypes and perform prototype-based personality inference using LLMs.

4.1 Data Augmentation and Prototype Building

We introduce an LLM-driven data augmentation pipeline prior to prototype construction (shown in Figure 2, see more in Appendix E, Algorithm 1). Let the labeled dataset be $U = \{\langle a_i, c_i \rangle\}_{i=1}^{n}$, where a_i denotes a user-generated text and $c_i \in \mathcal{C}^{\mathrm{MBTI}}$ is the corresponding personality label. We denote by $U' \subset U$ the subset used for data augmentation. We formalize the augmentation process

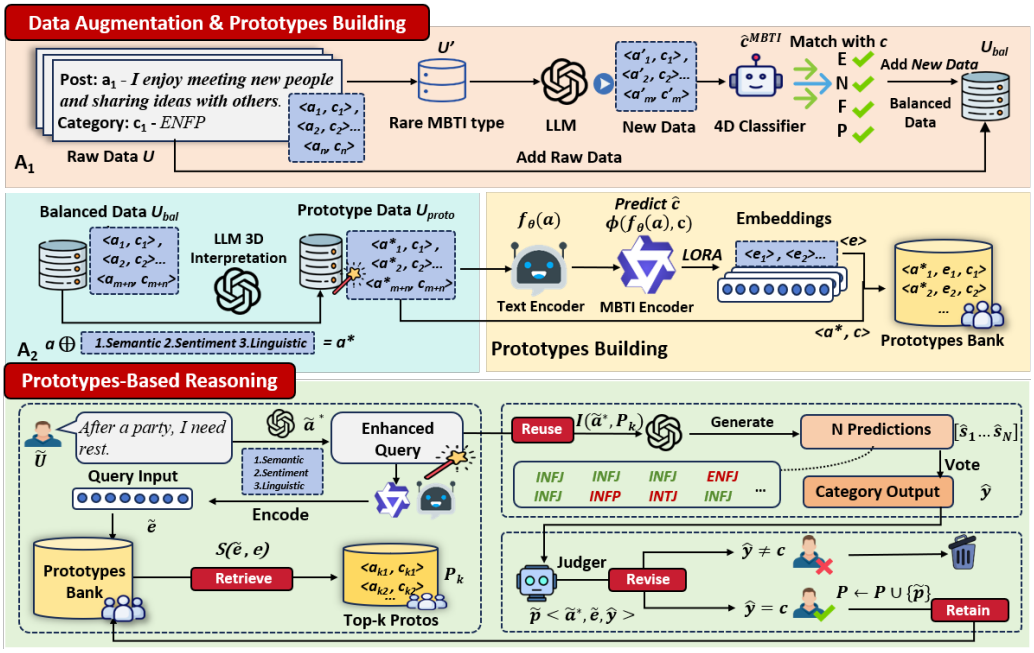

Figure 2: Overview of the data augmentation, prototype construction, and prototype-based reasoning framework. Rare MBTI categories are augmented using an LLM with few-shot prompting and filtered by a four-dimensional classifier, after which posts are interpreted along semantic, sentiment, and linguistic dimensions and stored in a structured prototype bank. Given a query, relevant prototypes are retrieved to generate candidate predictions, which are then refined through a judging process, with validated cases optionally retained for future reasoning.

as a composite operator $\mathcal{A} = \mathcal{A}_2 \circ \mathcal{A}_1$, where $\mathcal{A}_1 : U' \to U_{\text{bal}}$, $\mathcal{A}_2 : U_{\text{bal}} \to U_{\text{proto}}$. Here, $\mathcal{A}_1$ performs class-balanced augmentation with quality filtering, while $\mathcal{A}_2$ applies multi-dimensional representation expansion along semantic, linguistic, and affective dimensions.

Class-balanced augmentation with quality filtering ($\mathcal{A}_1$). To mitigate imbalance across MBTI categories, we first perform class-conditional data generation on samples in U'. Specifically, for any original sample $\langle a, c\rangle \in U'$, we guide the LLM using category-specific prompt templates (see Appendix G, Table 17) to generate a new text a' while preserving the original personality label, yielding a candidate sample $\langle a', c\rangle$. To ensure label consistency and usability, we introduce a 4D personality classifier(see Appendix C) as a quality filter. This classifier consists of a shared pretrained encoder, four binary classification heads corresponding to the MBTI dimensions (I/E, S/N, T/F, J/P), and an additional head for overall 16-type MBTI classification. Once trained to a performance level comparable to prior work (Lin et al., 2024), the classifier is fixed and used for filtering. A candidate sample $\langle a', c\rangle$ is accepted if and only if $\hat{c}^{\text{MBTI}} = c$ and $\hat{c}^{(i)} = c^{(i)}$, $\forall i \in \{1, 2, 3, 4\}$, where $\hat{c}^{\text{MBTI}}$ denotes the predicted MBTI type and $\hat{c}^{(i)}$ denotes the prediction for the i-th binary dimension. This process yields a class-balanced dataset U_{bal}.

Multi-dimensional representation augmentation ($\mathcal{A}_2$). After class-balanced augmentation, let $|U_{\text{bal}}| = n + m$, where n is the number of original samples and m is the number of newly generated samples. To further enrich prototype expressiveness, we apply the multi-dimensional augmentation operator $\mathcal{A}_2$ to *every* sample in U_{bal}. Concretely, guided by prompt templates defined in Appendix G, Table 4, $\mathcal{A}_2$ expands each text along semantic, linguistic, and affective dimensions without increasing the dataset size (Hu et al., 2024). For any sample $\langle a_j, c_j\rangle \in U_{\text{bal}}$, the augmented representation is defined as $a_j^* = \mathcal{A}_2(a_j)$. The resulting augmented dataset is $U_{\text{proto}} = \left\{\langle a_j^*, c_j\rangle \mid j = 1, 2, \ldots, n+m\right\}$, where a_j^* denotes the attribute-expanded text representation and $c_j \in \mathcal{C}^{\text{MBTI}}$ is the corresponding personality label.

Prototype Building. We fine-tune a compact encoder with LoRA on U_{proto} to learn *relation embeddings* between texts and personality categories. To avoid notational ambiguity, we decompose the representation learning process into two stages. First, a text encoder extracts a latent representation from the input text: $z = f_\theta(a) \in \mathbb{R}^d$. Then, a category-conditioned relation mapping

operator ϕ injects category information into the text representation, yielding a relation embedding: $e = \phi(z, c) = \phi(f_\theta(a), c) \in \mathbb{R}^d$. Accordingly, we formally define the relation embedding operator as $\mathcal{E} : (a, c) \mapsto e = \phi(f_\theta(a), c)$. During training, the model produces an overall MBTI type prediction $\hat{c} \in \mathcal{C}^{\text{MBTI}}$ via a 16-class classification head, and supervision is applied only at the fine-grained type level using the cross-entropy loss: $\mathcal{L}_{\text{proto}} = \text{CE}(\hat{c}, c)$. The resulting prototype set is denoted as $\mathcal{P} = \{\, p_j \mid j = 1, 2, \ldots, n + m \,\}$, where each prototype is constructed from an augmented sample $\langle a_j^*, c_j \rangle \in U_{\text{proto}}$ as $p_j = \langle a_j^*, e_j, c_j \rangle$, and stored in the prototype library (see Appendix E, Algorithm 3).

4.2 Prototype-Based Reasoning

Let the test set be $\tilde{U} = \{\tilde{a}\}$, which is disjoint from the training data and all augmented samples. At inference time, for any test instance $\tilde{a} \in \tilde{U}$, we first obtain its augmented representation $\tilde{a}^* = \mathcal{A}_2(\tilde{a})$ and extract a latent text representation using the same encoder: $\tilde{z} = f_\theta(\tilde{a}^*)$. We then obtain an initial MBTI type prediction $\hat{c} \in \mathcal{C}^{\text{MBTI}}$, which is used to construct a category-conditioned relation embedding: $\tilde{e} = \mathcal{E}(\tilde{a}^*, \hat{c}) = \phi(\tilde{z}, \hat{c}) \in \mathbb{R}^d$.

Retrieve. We define a similarity operator S using cosine similarity to measure the proximity between the test embedding $\tilde{e}$ and a prototype embedding e: $S(\tilde{e}, e) = \frac{\tilde{e} \cdot e}{\|\tilde{e}\| \|e\|}$. Based on this similarity measure, we define the prototype retrieval operator $\mathcal{R}_k : (\tilde{e}, \mathcal{P}) \mapsto \mathcal{P}_k$, where $\mathcal{P}_k = \{p_i\}_{i=1}^k$ denotes the set of the top-k prototypes most similar to $\tilde{e}$.

Reuse. As shown in Figure 2, given the retrieved prototype set $\mathcal{P}_k$, we define an LLM-based inference operator $\mathcal{I}$. Conditioning on the target text $\tilde{a}^*$ and its associated prototypes, $\mathcal{I}$ explicitly models the alignment between the test sample and historical prototypes via prompt templates (see Appendix G, Table 5), and outputs N predictive MBTI types: $\mathcal{I}(\tilde{a}^*, \mathcal{P}_k) \mapsto [\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_N]$, where each $\hat{s}_i \in \mathcal{C}^{\text{MBTI}}$ denotes a single predicted personality type, and duplicate predictions are allowed. The final predicted type $\hat{y}$ is obtained by majority voting over $[\hat{s}_i]_{i=1}^N$.

Revise & Retain. In evaluation or online inference scenarios where ground-truth labels c are available, if the prediction is correct, i.e., $\hat{y} = c$, the test sample is converted into a new prototype $\tilde{p} = \langle \tilde{a}^*, \tilde{e}, \hat{y} \rangle$, which is then added to the prototype set to support subsequent inference (see Appendix E, Algorithm 2): $\mathcal{P} \leftarrow \mathcal{P} \cup \{\tilde{p}\}$.

5 Experiments

Research Questions. To systematically evaluate the effectiveness of the ProtoMBTI framework and its cognitive plausibility, we formulate the following research questions:

- RQ1 (Section 6.1): *Does prototype-based personality detection achieve better performance than existing state-of-the-art (SOTA) models?*
- RQ2 (Section 6.2): *Can the effectiveness of the framework and the contribution of prototypes to the reasoning process be empirically validated?*
- RQ3 (Section 6.3): *Do prototypes preserve generalization ability across different test sets?*

Experimental Design. To address *RQ1*, we compare ProtoMBTI with multiple baseline methods on the Kaggle and Pandora datasets under the dichotomy-level setting. For *RQ2*, we conduct ablation studies to examine the contribution of prototypes to model performance and the reasoning process under different conditions. For *RQ3*, models are trained on a mixed training set constructed from Kaggle and Pandora, while the validation set and test set are drawn from different datasets to assess cross-dataset generalization.

Datasets and Evaluation Metrics. We use two standard MBTI datasets: Kaggle (8,675 samples from PersonalityCafe) and Pandora (9,067 samples from Reddit) (Gjurković et al., 2020). Both datasets are split into training, validation, and test sets with an 8:1:1 ratio, and the mixed training data is formed by directly merging the respective training sets. Data augmentation is applied only to the training and validation sets, while the test sets remain unchanged to ensure a fair evaluation of generalization. Model performance is evaluated using accuracy, including: (i) accuracy on the four MBTI dichotomies and their average (Hu et al., 2024; Bi et al., 2025), which is used for comparison with prior work; and (ii) fine-grained accuracy over all 16 MBTI types, serving as a more challenging metric for ablation studies and generalization evaluation. To ensure fair comparison, newly inferred samples are not incorporated into the prototype bank during inference. Additional experimental details are provided in Appendix D.

Model Configurations and Baselines. We select representative models for different components to

Methods	Kaggle					Pandora				
	I/E	S/N	T/F	P/J	Avg	I/E	S/N	T/F	P/J	Avg
$\text{Vanilla}_{\text{Qwen}}$	58.86	56.17	58.41	53.53	56.74	51.47	52.01	56.84	55.05	53.84
ICL_{Qwen}	50.23	44.31	47.86	48.19	47.65	46.18	47.57	53.30	52.26	49.83
TAE (Hu et al., 2024)	70.90	66.21	81.17	70.20	72.07	62.57	61.01	70.53	59.34	63.05
ETM (Bi et al., 2025)	68.97	71.21	86.19	84.78	77.79	68.57	64.91	66.07	63.53	65.77
$\text{ProtoMBTI}_{\text{llama}}$	81.92	87.70	86.04	82.47	84.03	69.05	68.85	68.98	70.82	69.43
$\text{ProtoMBTI}_{\text{Qwen}}$	83.74	88.10	84.54	84.18	85.14	71.63	66.98	73.25	70.33	70.55
$\text{ProtoMBTI}_{\text{GPT4o}}$	82.36	85.55	82.70	80.04	82.66	70.41	70.65	73.32	71.27	71.41

Table 1: Performance comparison of ProtoMBTI and LLM baselines on the Kaggle and Pandora datasets (see additional results in Appendix F). Metrics include the accuracies of four MBTI dimensions and their average, with subscripts denoting different backbone LLMs.

Method	Kaggle					
	I/E	S/N	T/F	P/J	Avg	16-T
ProtoMBTI-qwen	83.74	88.10	84.54	84.18	85.14	71.42
ProtoMBTI-Rand	81.54	83.69	80.62	70.77	79.66	50.77
ProtoMBTI-Zero	83.08	82.77	81.85	73.23	80.73	54.15
ProtoMBTI-2nd	80.92	82.77	80.92	78.15	80.69	59.08
ProtoMBTI-Sem	81.54	81.54	81.85	71.08	79.50	50.15
ProtoMBTI-Expl	82.77	89.45	78.66	73.58	81.12	56.62
ProtoMBTI-Raw	79.02	88.63	71.51	70.69	77.96	45.37
EncOnly-LLaMA	82.46	83.08	83.38	76.92	81.46	59.69
EncOnly-Qwen	80.00	83.69	83.38	76.62	80.92	60.62
EncOnly-DS	80.92	81.54	84.00	74.77	80.31	56.22

Table 2: Ablation results (accuracy, %) on Kaggle.

ProtoMBTI-Qwen								
Train	**Val**	**Test**	**I/E**	**S/N**	**T/F**	**P/J**	**Avg**	**16-T**
k+p	k	k	93.54	93.54	95.69	93.23	93.50	85.54
k+p	p	p	96.25	97.08	96.75	95.54	96.41	92.13
k+p	p	k	91.08	90.77	94.46	90.15	91.62	81.23
k+p	k	p	90.44	91.25	91.35	90.44	90.87	81.15

Table 3: Generalization results with mixed training data (Kaggle + Pandora). Validation and test sets are drawn from either Kaggle (k) or Pandora (p).

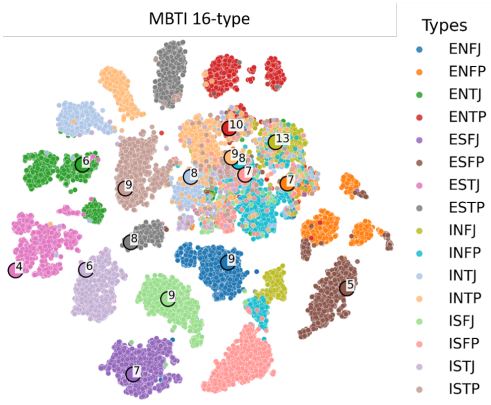

Figure 3: t-SNE visualization of the prototype bank on the Kaggle dataset. Each point represents the embedding of an MBTI type, with colors distinguishing personality categories. Large numbered circles denote prototypes most frequently retrieved during testing.

systematically evaluate ProtoMBTI under diverse configurations. For post generation and interpretation, we use GPT-4o and GPT-4o-mini to compare the effects of different model scales. In the data augmentation stage, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2020) are employed as backbone models for the four-dimensional classifier to control the quality of LLM-generated text. For feature extraction, we adopt DeepSeek-1B (Bi et al., 2024), Qwen2.5-1.5B (Bai et al., 2025), and Llama3-1B (Dubey et al., 2024) to examine the impact of encoder architecture and scale on personality representation learning. During inference, we compare GPT-4o-mini, Qwen2-72B (Bai et al., 2025), and Llama3.1-70B (Dubey et al., 2024) to evaluate the influence of different reasoning engines on final predictions.

In the main text, we report comparisons between ProtoMBTI and recent LLM-based SOTA methods based on Qwen2-72B, including both vanilla and in-context learning (ICL) (Dong et al., 2024) settings, TAE (Hu et al., 2024), and ETM (Bi et al., 2025). Comprehensive comparisons with all baselines and detailed prompts are provided in Appendix H.

Implementation and Experimental Environment. All experiments are implemented using PyTorch (Paszke et al., 2019) and Huggingface Transformers (Wolf et al., 2019). Model training is conducted on NVIDIA A100 and RTX 4090 GPUs, while large-scale inference relies on official APIs. Random seeds are fixed throughout all experiments to guarantee result stability. Further implementation and training details are available in Appendix D.

6 Results and Discussion

Overall, we address RQ1 through performance comparisons, RQ2 through ablation studies, and

Figure 4: Comparison of ISTP personality traits (left) and ProtoMBTI model outputs (right). Highlighted words indicate cues aligned with ISTP attributes.

RQ3 through mixed-data generalization experiments. All reported results correspond to averaged performance across the evaluated datasets and experimental settings. In addition, we derive psychological insights from the experimental results and conduct a case study. Discussions on data augmentation, hyper-parameters, and additional experimental results are provided in Appendix I.

6.1 Performance Comparison (RQ1)

As shown in Table 1, the proposed ProtoMBTI framework consistently outperforms existing methods under the evaluation metrics adopted in prior work. On the four MBTI dichotomies, ProtoMBTI$_{\text{Qwen}}$ achieves an average accuracy of 85.14% on the Kaggle dataset, substantially surpassing the previously best-performing model ETM (77.79%). On the Pandora dataset, ProtoMBTI$_{\text{GPT4o}}$ reaches 71.41%, again exceeding ETM (65.77%). These results indicate that, under single-dataset settings, ProtoMBTI outperforms current state-of-the-art methods on the four-dimension classification task.

6.2 Ablation Study (RQ2)

Table 2 presents the ablation results on the Kaggle dataset. These results highlight the central roles of prototypes, data augmentation, and prototype-based reasoning in the model. We take ProtoMBTI$_{\text{Qwen}}$ as the full model and compare it with its ablated variants on the Kaggle dataset.

First, effective prototype selection is crucial: when only suboptimal prototypes are used (ProtoMBTI$_{\text{2nd}}$), the 16-type accuracy drops by 12.34 percentage points compared to the full model (from 71.42% to 59.08%). When random prototypes (ProtoMBTI$_{\text{Rand}}$) or simple semantic retrieval (ProtoMBTI$_{\text{Sem}}$) are employed, the average accuracy over the four dichotomies decreases by 5.48 and 5.64 percentage points, respectively, while the 16-type accuracy drops by 20.65 and 21.27 percentage points. Performance under these settings is even worse than that of ProtoMBTI$_{\text{Zero}}$, which does not use any prototypes and exhibits a 17.27 percentage-point decrease in 16-type accuracy relative to the full model. These results suggest that prototype selection directly determines model performance: inappropriate prototypes not only fail to provide useful information but can also substantially interfere with the reasoning process.

Second, removing category balancing or explanation-based data augmentation also significantly degrades performance: ProtoMBTI$_{\text{Expl}}$, which retains only explanation-based augmentation, shows a 14.80 percentage-point drop in 16-type accuracy, while ProtoMBTI$_{\text{Raw}}$, which removes augmentation entirely, suffers a larger drop of 26.05 percentage points. This indicates that high-quality and well-controlled data augmentation is critical for effective prototype construction.

Third, removing prototype-based reasoning leads to the most severe performance degradation. Encoder-only models achieve a maximum 16-type accuracy of 60.62%, corresponding to a 10.80 percentage-point decrease relative to ProtoMBTI$_{\text{Qwen}}$, and remain substantially inferior to the full model with prototype-based reasoning.

Moreover, the 16-type metric is more sensitive than the average accuracy over the four

dichotomies: the performance gap between ProtoMBTI$_{\text{Qwen}}$ and ProtoMBTI$_{\text{Raw}}$ reaches 26.05 percentage points on the 16-type task, whereas the average accuracy over the four dichotomies decreases by only 7.18 percentage points. This further demonstrates that fine-grained personality classification relies more heavily on model structure and reasoning mechanisms.

6.3 Generalization Experiments (RQ3)

When trained on mixed data from Kaggle and Pandora(Table 3), with validation and test sets drawn from the same dataset, ProtoMBTI consistently outperforms single-dataset training across all metrics, indicating that mixed training facilitates the learning of more generalizable representations. On the 16-type classification task, the model achieves 85.54% on Kaggle and 92.13% on Pandora.

Under the cross-dataset setting, where validation and test sets come from different datasets, ProtoMBTI still maintains stable performance: the 16-type accuracy reaches 81.23% on Kaggle and 81.15% on Pandora, while the average accuracy over the four dichotomies reaches 91.62% and 90.87%, respectively. Although performance slightly decreases compared to the same-domain setting, it remains substantially better than single-dataset training, demonstrating strong cross-dataset generalization.

6.4 Psychological Analysis

Prototype theory suggests that human category judgments typically rely on a small number of representative exemplars and proceed through increasingly fine-grained abstraction levels, rather than being based on rigid and sharply defined category boundaries (Rosch, 1973, 1975; Rosch et al., 1976). As a result, typical members are more easily recognized, higher-level category judgments tend to be more stable, and category membership often exhibits continuity and overlap.

Our experimental results align with these cognitive characteristics in several respects. First, as shown in Table 1, ProtoMBTI prioritizes representative prototypes rather than averaging over all samples, leading to consistent improvements over existing methods across all metrics. The ablation results further indicate that only highly representative prototypes contribute positively to performance, whereas random or non-typical prototypes tend to interfere with inference and even degrade performance. Second, under mixed-training and cross-dataset settings, the model maintains stable performance, suggesting that it captures patterns shared across data sources rather than relying on dataset-specific surface features.

In addition, predictions on the four MBTI dichotomies are consistently more stable than fine-grained 16-type classification, particularly in ablation and generalization settings. This observation suggests that higher-level categories are more reliably distinguished, consistent with the cognitive tendency to reason at coarser levels of abstraction.

Finally, the prototype distributions shown in Figure 3 exhibit both clustering tendencies and overlaps among different MBTI types. This finding is consistent with prior psychological studies indicating that personality type boundaries are not strictly separable (Stein and Swan, 2019; Capraro and Capraro, 2002; Erford et al., 2025), and suggests that prototype distributions capture not only clustering structure but also cognitive confusability across categories.

6.5 Case Study

Figure 4 presents a case from the real test set together with the prototypes retrieved from the prototype bank. The left panel shows the official MBTI description of the ISTP type, characterized by pragmatism, rationality, problem-solving orientation, and restrained emotional expression. The right panel illustrates ProtoMBTI's reasoning process on a social media post: expressions such as "cut the noise," "fix problems," and "don't waste time whining or explaining" reflect a direct and pragmatic stance; sentiment analysis indicates determination and a sense of control, while linguistic style analysis reveals concise, forceful, and emotionally restrained language. The high-frequency retrieved prototypes further emphasize patterns such as prioritizing action over words and avoiding unnecessary explanations, leading the model to classify the post as ISTP, in close agreement with the official MBTI description.

7 Conclusion

We introduce ProtoMBTI, a prototype-based framework for MBTI personality detection. By grounding inference in prototype theory, ProtoMBTI achieves robust performance and improves interpretability, demonstrating the potential of cognitively motivated modeling for personality inference.

8 Limitations

Despite the strong empirical performance of ProtoMBTI, several limitations remain. For details regarding the usage of LLMs, please refer to Appendix A.

First, while LLM-based data augmentation plays a crucial role in improving performance, we fix prompt templates and generation protocols to ensure reproducibility and release these details for verification. Although the proposed framework is not tied to a specific backbone, different choices of LLMs may lead to performance variations, which are not exhaustively explored in this work.

Second, although prior studies have not directly reported 16-type MBTI classification accuracy, and ProtoMBTI substantially outperforms existing methods under both average and dimension-level metrics, future work should incorporate stronger direct multi-class baselines to further strengthen empirical comparisons.

Third, we view ProtoMBTI as a promising step rather than a final solution. Broader validation across additional datasets, domains, and usage scenarios is required to assess its robustness beyond the benchmarks considered in this study.

Fourth, the theoretical grounding of this work primarily draws on classical prototype theory. While this provides a cognitively motivated foundation, future studies may integrate recent advances in computational cognitive science to further extend and refine the proposed framework.

Fifth, although the datasets used in this study cover the major publicly available MBTI text resources, they are limited to a single language. Consequently, the applicability and robustness of ProtoMBTI in multilingual and cross-cultural settings remain unexplored, which is particularly relevant given the ACL community's emphasis on multilingual evaluation.

9 Ethical Considerations

ProtoMBTI is designed to model MBTI-based personality categories as an operational and descriptive framework rather than as a clinical or psychological diagnostic tool. Predictions produced by ProtoMBTI reflect probabilistic patterns in language use and should not be interpreted as definitive, authoritative, or prescriptive assessments of an individual's personality.

As with prior work in computational personality analysis, there is a risk of overinterpretation or misuse if such models are applied in high-stakes or consequential settings, such as hiring, screening, or psychological evaluation. We therefore caution against deploying ProtoMBTI in decision-making scenarios that may materially affect individuals without appropriate human oversight, domain expertise, and contextual understanding. ProtoMBTI and its associated artifacts are intended exclusively for research and exploratory analysis of personality-related patterns in text, particularly for advancing methodological understanding in computational modeling and cognitively grounded AI systems.

All datasets used in this work are publicly available and were released with prior anonymization and content moderation by their original creators. ProtoMBTI does not introduce, infer, or reconstruct personally identifying information, nor does it attempt to associate textual data with real-world identities. While the source datasets may contain naturally occurring subjective opinions or emotionally charged language, this work does not curate, amplify, or generate offensive content beyond what is already present in the original data sources. Nevertheless, responsible use of ProtoMBTI requires awareness of broader ethical considerations related to user consent, fairness, and potential downstream societal impacts.

All artifacts associated with ProtoMBTI, including source code, prompts, and configuration files, are released under a permissive open-source license (e.g., MIT License). The license permits use, modification, and redistribution for research and non-commercial purposes, while disclaiming warranties and limiting liability. Users are responsible for ensuring that any downstream use of the artifacts complies with applicable legal, ethical, and institutional requirements and remains consistent with the intended scope of the framework.

To promote transparency, reproducibility, and responsible reuse, we provide comprehensive documentation covering model architecture, training procedures, prompt design, and evaluation protocols. The anonymized codebase and accompanying documentation are publicly available at:

https://anonymous.4open.science/r/ProtoMBTI-6F00

References

Agnar Aamodt and Enric Plaza. 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, 7(1):39–59.

Nimra Ashraf, Rao Sohail Ahmad, Shehar Bano, Hafiz Muhammad Azeem, and Shagufta Naz. 2024. Enhancing mbti personality prediction from text data with advance word embedding technique. *VFAST Transactions on Software Engineering*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Weihong Bi, Feifei Kou, Lei Shi, Yawen Li, Haisheng Li, Jinpeng Chen, and Mingying Xu. 2025. Leveraging the dual capabilities of llm: Llm-enhanced text mapping model for personality detection. In *AAAI Conference on Artificial Intelligence*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Robert M Capraro and Mary Margaret Capraro. 2002. Myers-briggs type indicator score reliability across: Studies a meta-analytic reliability generalization study. *Educational and Psychological Measurement*, 62(4):590–602.

Brandon Cui and Calvin Qi. 2017. Survey analysis of machine learning methods for natural language processing for mbti personality type prediction. *Final Report Stanford University*.

Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Neuron-based personality trait induction in large language models. *arXiv preprint arXiv:2410.12327*.

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.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128.

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

Bradley T Erford, Xi Zhang, Elizabeth Sweeting, Mia Russo, Anna Rashid, Martin F Sherman, Emily L Bradford, Xinran Wang, Allison Gao, Xinlei Huang, and 1 others. 2025. A 25-year review and psychometric synthesis of the myers–briggs type indicator (mbti)–form m. *Journal of Counseling & Development*.

Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Šnajder. 2020. Pandora talks: Personality and demographics on reddit. *arXiv preprint arXiv:2004.04460*.

Kostas Hatalis, Despina Christou, and Vyshnavi Kondapalli. 2025. Review of case-based reasoning for llm agents: theoretical foundations, architectural components, and cognitive integration. *arXiv preprint arXiv:2504.06943*.

Feng He, Zijun Chen, Xinnian Liang, Tingting Ma, Yunqi Qiu, Shuangzhi Wu, and Junchi Yan. 2025. Protoreasoning: Prototypes as the foundation for generalizable reasoning in llms. *arXiv preprint arXiv:2506.15211*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial personalized ranking for recommendation. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.

Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 2024. Llm vs small model? large language model based text augmentation enhanced personality detection model. In *AAAI Conference on Artificial Intelligence*.

Hang Jiang, Xianzhe Zhang, and Jinho D. Choi. 2020. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence: Student Abstract and Poster Program*, pages 13821–13822.

Oliver P. John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives.

Sedrick Scott Keh, I Cheng, and 1 others. 2019. Myers-briggs personality classification and personality-specific language generation using pre-trained language models. *arXiv preprint arXiv:1907.06333*.

Janet L Kolodner. 1992. An introduction to case-based reasoning. *Artificial intelligence review*, 6(1):3–34.

Mike Komisin and Curry I. Guinn. 2011. Identifying personality types using document classification methods. In *The Florida AI Research Society*.

Bohan Li, Jiannan Guan, Longxu Dou, ylfeng, Dingzirui Wang, Yang Xu, Enbo Wang, Qiguang Chen, Bichen Wang, Xiao Xu, Yimeng Zhang, Libo Qin, Yanyan Zhao, Qingfu Zhu, and Wanxiang Che. 2025. Can large language models understand you better? an mbti personality detection dataset aligned with population traits. In *International Conference on Computational Linguistics*.

Jiayu Li, Peijie Sun, Zhefan Wang, Weizhi Ma, Y. Li, M. Zhang, Zhoutian Feng, and Daiyue Xue. 2023. Intent-aware ranking ensemble for personalized recommendation. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Pei-Lun Li, Xiaomeng Liu, and Yongxing Wang. 2024. A novel method based on large language model for mbti classification: A novel mbti classification method. *Proceedings of the 2024 International Conference on Computer and Multimedia Technology*.

I-Fan Lin, Faegheh Hasibi, and Suzan Verberne. 2024. Generate then refine: data augmentation for zero-shot intent detection. *arXiv preprint arXiv:2410.01953*.

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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Veronica Lynn, Niranjan Balasubramanian, and H. Andrew Schwartz. 2020. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5306–5316.

Xingkong Ma, Houjie Qiu, Shujia Yao, Xinyi Chen, Jingsong Zhang, Zhaoyun Ding, Shaoyong Li, and Bo Liu. 2022. A general personality analysis model based on social posts and links. In *Pacific Rim International Conference on Artificial Intelligence*, pages 289–303. Springer.

Robert R. McCrae and Paul T. Costa. 1989. Reinterpreting the myers-briggs type indicator from the perspective of the five-factor model of personality. *Journal of personality*, 57 1:17–40.

Isabel Briggs Myers. 1962. The myers-briggs type indicator.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *ArXiv*, abs/2307.16180.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Suman A. Patil, Shivleela Patil, and Vijayalaxmi V. Tadkal. 2024. Enhanced personality prediction using knowledge distillation with bert: A focus on mbti. *Optical Memory and Neural Networks*.

Naomi L. Quenk. 1999. Essentials of myers-briggs type indicator assessment.

Zhaochun Ren, Zhou Yang, Chenglong Ye, Yufeng Wang, Haizhou Sun, Chao Chen, Xiaofei Zhu, Yunbing Wu, and Xiangwen Liao. 2024. E-icl: Enhancing fine-grained emotion recognition through the lens of prototype theory. *arXiv preprint arXiv:2406.02642*.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

Eleanor Rosch and Carolyn B Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439.

Eleanor H Rosch. 1973. Natural categories. *Cognitive psychology*, 4(3):328–350.

Gregorius Ryan, Pricillia Katarina, and Derwin Suhartono. 2023. Mbti personality prediction using machine learning and smote for balancing data based on statement sentences. *Inf.*, 14:217.

Ramteja Sajja, Yusuf Sermet, Muhammed Cikmaz, David Cwiertny, and Ibrahim Demir. 2023. Artificial intelligence-enabled intelligent assistant for personalized and adaptive learning in higher education. *ArXiv*, abs/2309.10892.

Aditya G Shanmukha, R.S Shamyuktha, S Karan, Deepa Gupta, and Suja Palaniswamy. 2024. Advancing personality detection through word embedments and deep learning: An examination using the mbti dataset. *2024 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pages 1–6.

Dr. V. Shobha, Rani Asst.Professor, Dr. A. Ramesh Babu, Chittireddy Akhil Reddy, Vallem Randheer, Reddy Asst.Professor, Dr.V. Ramu, Asst. Professor, and B. Shruthi. 2024. Mbti personality type prediction using bert-lstm and deep learning on social media posts. *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, pages 984–989.

Randy Stein and Alexander B Swan. 2019. Evaluating the validity of myers-briggs type indicator theory: A teaching tool and window into intuitive psychology. *Social and Personality Psychology Compass*, 13(2):e12434.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly D. Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307.

Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2018. Personality predictions based on user behavior on the facebook social media platform. *IEEE Access*, 6:61959–61969.

Tommy Tandera, Derwin Suhartono, Rini Wongso, Yen Lina Prasetio, and 1 others. 2017. Personality prediction system from facebook users. *Procedia computer science*, 116:604–611.

Raad Bin Tareaf. 2022. Mbti bert: A transformer-based machine learning approach using mbti model for textual inputs. In *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, pages 2285–2292.

Yue Wang. 2024. Ai and mbti: A synergistic framework for enhanced team dynamics. *ArXiv*, abs/2409.15293.

Bowen Wei, Mehrdad Fazli, and Ziwei Zhu. 2025. Learning to explain: Prototype-based surrogate models for llm classification. *arXiv preprint arXiv:2505.18970*.

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering. In *International Conference on Case-Based Reasoning*, pages 445–460. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Di Xue, Lifa Wu, Zheng Hong, Shize Guo, Liang Gao, Zhiyong Wu, Xiaofeng Zhong, and Jianshan Sun. 2018. Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*, 48(11):4232–4246.

Feifan Yang, Xiaojun Quan, Yunyi Yang, and Jianxing Yu. 2021a. Multi-document transformer for personality detection. In *AAAI Conference on Artificial Intelligence*.

Tao Yang, Jinghao Deng, Xiaojun Quan, and Qifan Wang. 2022. Orders are unwanted: Dynamic deep graph convolutional network for personality detection. In *AAAI Conference on Artificial Intelligence*.

Tao Yang, Jinghao Deng, Xiaojun Quan, and Qifan Wang. 2023. Orders are unwanted: dynamic deep graph convolutional network for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13896–13904.

Tao Yang, Feifan Yang, Haolan Ouyang, and Xiaojun Quan. 2021b. Psycholinguistic tripartite graph network for personality detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4229–4239.

Hanwen Zhang. 2023. Mbti personality prediction based on bert classification. *Highlights in Science, Engineering and Technology*.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2024a. Personality alignment of large language models. *arXiv preprint arXiv:2408.11779*.

Yangfu Zhu, Yue Xia, Meiling Li, Tingting Zhang, and Bin Wu. 2024b. Data augmented graph neural networks for personality detection. In *AAAI Conference on Artificial Intelligence*.

A The Use of Large Language Models(LLMs)

In this work, Large Language Models (LLMs) were used in the following auxiliary capacities:

1. **Data augmentation:** Prompt templates for generating synthetic posts were drafted with the assistance of GPT-4o, ensuring linguistic diversity and alignment with MBTI personality categories. The final prompts were manually verified and refined by the authors.

2. **Code generation:** Portions of the experimental codebase were initially drafted using an editor equipped with an LLM assistant (based on GPT-4.1). These drafts were strictly treated as scaffolding; all implementations were subsequently checked, rewritten where necessary, and validated by the authors to guarantee correctness and reproducibility.

3. **Manuscript refinement:** GPT-5 was employed for polishing the writing, including grammar correction, wording suggestions, and restructuring of some paragraphs. Importantly, the intellectual contributions—including research design, theoretical framing, dataset construction, experiments, and analyses—were carried out entirely by the authors.

4. **Dataset handling:** All datasets used in this study are publicly available (Kaggle and Pandora MBTI corpora). Prior to any use with LLMs, we performed preprocessing and cleaning to ensure that no sensitive or personally identifiable information (PII) was input into the models.

All other aspects of this study—including literature review, methodological design, data processing, model training, evaluation, interpretation of results, and theoretical grounding—were performed solely by the human authors. The LLMs served only as auxiliary tools to improve efficiency and clarity; they did not contribute to the conceptual novelty or scientific insights of this work.

B Prototype Theory Insights

Details on Prototype Construction and Reasoning. Prototype construction aligns with the *Prototype Effect* in psychology: prototypes are abstracted from long-term experience and serve as cognitive anchors that represent the most typical members of a category. In our setting, Kaggle and Pandora posts are regarded as accumulated experiential data, and semantic embeddings are trained to internalize these experiences. To ensure psychological plausibility, we balance sample distributions across MBTI categories and apply quality filtering, so that embeddings faithfully represent their categories rather than spurious artifacts. This construction ensures that frequently invoked prototypes occupy central positions within clusters, mirroring the notion that typical members are more cognitively salient than atypical ones.

Beyond the prototype effect, the construction also reflects *graded membership*: within each MBTI type, some posts are more representative than others, and our selection strategy assigns higher weight to prototypes that are more frequently retrieved during inference. This graded salience ensures that the prototype bank does not treat all members as equal, but rather reflects the natural hierarchy of typicality within categories.

Prototype-based reasoning follows a *retrieve–reuse–revise–retain* cycle: new inputs are matched against existing prototypes, adapted through linguistic and behavioral cues, corrected if inconsistent, and retained if verified. This cycle parallels Case-Based Reasoning (CBR) (Hatalis et al., 2025; Wiratunga et al., 2024; Aamodt and Plaza, 1994; Kolodner, 1992), but differs by explicitly modeling the cognitive internalization process suggested by prototype theory. Specifically, retrieval captures the *family resemblance* principle: inputs are not compared on rigid boundaries but by overlapping features with prototypes, reflecting the fuzzy nature of MBTI category boundaries observed in psychology.

Finally, the dual-level supervision design (dimensions vs. types) embodies the notion of *basic-level categories*. The four MBTI dichotomies act as higher-level categories, while the 16 MBTI types correspond to finer-grained subcategories. By grounding inference at both levels, ProtoMBTI captures the cognitive process of transitioning smoothly from coarse categories to specific exemplars. This hierarchical organization reflects the graded structure between superordinate, basic, and subordinate levels emphasized in prototype theory.

In summary, prototype construction operationalizes the *prototype effect* and *graded membership*, while prototype-based reasoning integrates *family resemblance* and *basic-level categories*. Thus, ProtoMBTI does not merely apply prototypes as a computational trick, but instantiates them as cognitively grounded mechanisms of categorization.

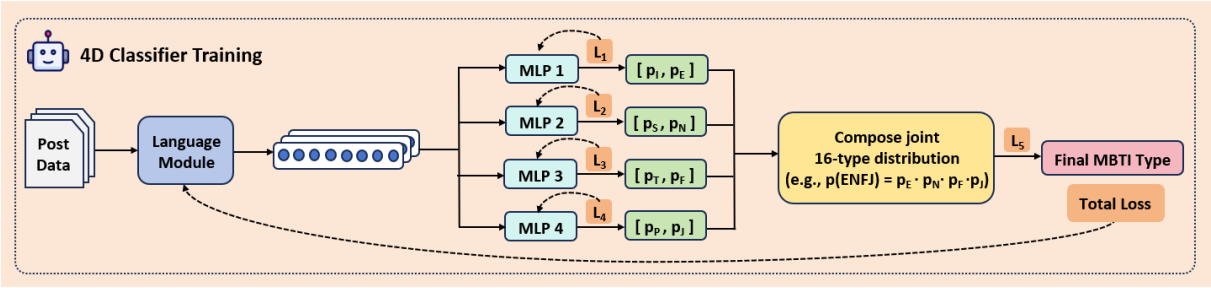

Figure 5: Overview of the 4D classifier training framework for MBTI prediction. User post data are first encoded by a shared language module to obtain a unified textual representation. This representation is then fed into four parallel MLP classifiers, each corresponding to one MBTI dimension: Extraversion–Introversion (E/I), Sensing–Intuition (S/N), Thinking–Feeling (T/F), and Judging–Perceiving (P/J). Each MLP produces a binary probability distribution for its respective dimension and is supervised with an individual loss. The four dimension-wise probabilities are subsequently composed into a joint 16-type MBTI distribution via factorized combination (e.g., $p(\text{ENFJ}) = p_E \cdot p_N \cdot p_F \cdot p_J$). A final loss is applied on the composed distribution, and all loss terms are jointly optimized in an end-to-end manner, enabling simultaneous learning of dimension-level discrimination and coherent type-level inference.

C 4D Classifier

Training setup. The raw data are split into training, validation, and test sets with an 8:1:1 ratio, see in figure 5. Formally, given an input post x, an encoder f_θ maps the text into a latent representation. This representation is shared across multiple prediction heads: four dichotomy heads $g_{\phi_i}^{(i)}, i \in \{1,2,3,4\}$ corresponding to the four MBTI dimensions, and one type-level head $g_\psi^{(\text{type})}$ corresponding to the full 16-type classification. The predictions are given by $\hat{c}^{(i)} = g_{\phi_i}^{(i)}(f_\theta(x)), \quad \hat{c}^{\text{type}} = g_\psi^{(\text{type})}(f_\theta(x))$.

Loss functions. Let $c^{(i)}$ denote the ground-truth label of the i-th MBTI dimension and c the ground-truth 16-type label. We employ standard cross-entropy loss for both dimension-level and type-level tasks: $\mathcal{L}_{\text{dim}} = \frac{1}{4}\sum_{i=1}^{4} \text{CE}(\hat{c}^{(i)}, c^{(i)}), \quad \mathcal{L}_{\text{type}} = \text{CE}(\hat{c}^{\text{type}}, c)$.

The dimension loss $\mathcal{L}_{\text{dim}}$ encourages correct classification across the four dichotomies, while the type loss $\mathcal{L}_{\text{type}}$ directly supervises the fine-grained 16-type prediction.

Gradient updates. During training, each dichotomy head $g_{\phi_i}^{(i)}$ is updated with gradients from its own cross-entropy loss $\nabla\text{CE}(\hat{c}^{(i)}, c^{(i)})$, and the type-level head $g_\psi^{(\text{type})}$ is updated by $\nabla\mathcal{L}_{\text{type}}$. The encoder f_θ is updated by a balanced combination of both supervision signals: $\mathcal{L}_{\text{enc}} = \frac{1}{2}\big(\mathcal{L}_{\text{type}} + \mathcal{L}_{\text{dim}}\big)$.

This design ensures that the encoder simultaneously learns to capture broad dichotomy-level information and fine-grained 16-type discriminative features. In practice, this joint optimization stabilizes training and improves generalization across datasets.

Rationale for joint supervision. The use of both dimension-level and type-level supervision is motivated by the cognitive principle of *basic-level categories* in prototype theory (Rosch, 1975). In human categorization, individuals tend to reason at an intermediate level of abstraction: basic-level categories (e.g., "chair") are cognitively more salient than superordinate categories (e.g., "furniture") or subordinate categories (e.g., "rocking chair").

In the MBTI setting, the four dichotomies (I/E, S/N, T/F, J/P) can be viewed as higher-level dimensions, whereas the 16 types represent finer-grained subcategories. By jointly supervising the encoder with both dimension-level and type-level signals, ProtoMBTI encourages representations that are consistent across levels of categorization. This allows the encoder to learn (i) robust general features that align with dichotomous personality dimensions, and (ii) discriminative features necessary for fine-grained type prediction.

From a modeling perspective, this joint training mitigates the risk of overfitting to either overly coarse (dimension-only) or overly fine-grained (type-only) supervision. From a cognitive perspective, it operationalizes the graded relationship between superordinate, basic-level, and subordinate categories as described in prototype theory, ensuring that the learned prototypes function as psychologically plausible category exemplars.

D Experiment Setup

Detailed Experimental Design. For *RQ1*, we design *Main Experiment 1*, comparing ProtoMBTI and baseline models on Kaggle and Pandora in both the four MBTI dichotomies (I/E, S/N, T/F, J/P) and the full 16-type classification. For *RQ2*, we run a series of ablation studies to isolate the role of prototypes in reasoning. The conditions include: (i) top-k prototype retrieval; (ii) interval-based retrieval ($[k+1, 2k]$); (iii) random prototype selection; (iv) no prototypes, with only multi-dimensional explanation of raw data; and (v) no data augmentation. For *RQ3*, we conduct cross-domain transfer experiments by training on mixed datasets while validating on a single source, and by evaluating transfer between Kaggle and Pandora in both directions. We analyze performance degradation in both dichotomy-level and 16-type classification to assess generalization under distribution shift.

Dataset Details. The Kaggle dataset consists of 8,675 users, each with a four-letter MBTI type and excerpts from their 50 most recent posts. The Pandora dataset comprises 9,067 Reddit users, offering a more diverse linguistic distribution. Detailed pre- and post-augmentation distributions across MBTI types and dimensions are shown in Tables 6, 7, 8, and dataset splits are listed in Table 9. Only training and validation sets undergo augmentation; test sets remain original.

Metric Details. For comparability, we adopt accuracy as the main metric. At the higher level, accuracy is reported for each MBTI dichotomy (I/E, S/N, T/F, J/P) and their average. At the finer level, we report accuracy over all 16 MBTI types, which offers a more comprehensive measure of model performance. Since prior studies did not directly report 16-type performance, we compute theoretical results by multiplying the four dichotomy accuracies under MBTI logic.

Implementation Details. We use PyTorch (Paszke et al., 2019) with Huggingface Transformers (Wolf et al., 2019) for all implementations. Optimization follows AdamW (Loshchilov and Hutter, 2017) with an initial learning rate of 2×10^{-5}, batch size of 32, and 10 epochs. Experiments are run on an NVIDIA A100 GPU (80GB) and, for smaller-scale runs, on an NVIDIA RTX 4090. For large-scale inference with models exceeding local GPU memory, we rely on official APIs. All experiments are conducted with fixed random seeds to guarantee result stability.

E Algorithm

Algorithm 1 LLM-Driven Data Augmentation

Require: Original labeled dataset $U = \{\langle a, c\rangle\}$; class set $\mathcal{C}^{\mathrm{MBTI}}$; subset $U' \subset U$; LLM with prompt templates (see Appendix); a trained 4D Classifier as gatekeeper

Ensure: Augmented dataset U_{proto}

Stage A1: Class-balanced augmentation with quality filtering

for all class $c \in \mathcal{C}^{\mathrm{MBTI}}$ **do**
  Determine target count to balance class c
  while class c is under target **do**
    Select seed $\langle a, c\rangle$ from U' (or U)
    Use LLM + class-specific prompt to generate candidate $\langle a', c\rangle$
    Run 4D Classifier on a' to obtain predicted type and four dichotomies
    if predicted type $= c$ **and** each predicted dichotomy matches c **then**
      Accept $\langle a', c\rangle$ into $\mathcal{A}_1(U')$
    end if
  end while
end for
Form $U^{(1)} \leftarrow U \cup \mathcal{A}_1(U')$

Stage A2: Multi-dimensional augmentation

for all $\langle a, c\rangle \in U^{(1)}$ **do**
  Apply LLM-based *semantic* augmentation to obtain variant(s)
  Apply LLM-based *linguistic* augmentation to obtain variant(s)
  Apply LLM-based *sentiment* augmentation to obtain variant(s)
  Merge attribute-extended representation(s) into a^*
end for
Assemble $U_{\mathrm{proto}} \leftarrow \{\langle a^*, c\rangle \mid \langle a, c\rangle \in U^{(1)}\}$
Optional: run a final quality-control pass on generated items; remove low-quality samples
return U_{proto}

Analysis of Algorithm 1. The augmentation algorithm proceeds in two major stages designed to address distinct challenges in MBTI text classification. Stage A1 targets *class imbalance* by iteratively generating synthetic samples for under-represented categories. Instead of blindly trusting

LLM outputs, a dedicated 4D Classifier acts as a gatekeeper to ensure label fidelity at both the dichotomy and full-type levels. This filtering step is essential to prevent label noise, which would otherwise dilute the quality of the prototype bank.

Stage A2 enriches the representational space by applying *multi-dimensional augmentations* (semantic, linguistic, sentiment) to all available samples. Rather than expanding the dataset size indefinitely, each instance is transformed into an attribute-extended representation, ensuring diversity of expression without inflating sample counts. This design maintains computational efficiency while increasing robustness to stylistic and affective variability in real-world posts.

Overall, the algorithm ensures that the final augmented dataset U_{proto} achieves three desirable properties: (i) **balanced distribution** across MBTI types, (ii) **high fidelity** through classifier-verified filtering, and (iii) **rich expressiveness** via controlled augmentation dimensions. These characteristics jointly improve the stability of prototype construction and inference, especially under cross-domain distribution shifts.

Analysis of Algorithm 3. The prototype construction procedure transforms the augmented dataset U_{proto} into a structured prototype bank $\mathcal{P}$. The process begins by fine-tuning a compact LLM encoder f_θ using LoRA. This choice balances two competing objectives: (i) sufficient capacity to capture MBTI-specific textual nuances, and (ii) computational efficiency compared to large-scale models. Training is supervised at the 16-type classification level, ensuring embeddings reflect personality distinctions at the most granular MBTI resolution.

Each sample is then mapped to a prototype triplet $\langle a, r, c \rangle$, where r denotes the semantic embedding aligned with label c. Unlike traditional label-only storage, this triplet representation preserves both the linguistic surface form (a) and its learned relational embedding (r), enabling exemplar-driven retrieval during inference. Optional organization by class further facilitates efficient prototype access.

Overall, Algorithm 3 ensures that the resulting prototype bank has three desirable properties: (i) **discriminative power**, since embeddings are trained with supervised MBTI signals; (ii) **interpretability**, as each prototype links a real text instance to its embedding and type; and (iii) **extensibility**, allowing incremental updates as new verified cases are added during inference. These properties make $\mathcal{P}$ a cognitively plausible and computationally tractable foundation for prototype-driven reasoning.

F Additional Experiments

Analysis of Algorithm 2. The inference procedure integrates prototype retrieval with LLM-based reasoning to align prediction with psychological intuition. Each unseen post is first augmented (A_2) to enrich stylistic and semantic variability, ensuring that inference is not overly sensitive to surface-level expression. The post is then encoded into an embedding r' via the inference encoder $\mathcal{E}^*$ and compared against the prototype bank $\mathcal{P}$ using cosine similarity. This design enables inference by *analogy*, where predictions are grounded in similarity to previously observed exemplars.

The operator $\mathcal{I}$ incorporates both the target post and the retrieved prototypes into a prompt, guiding the LLM to perform case-based reasoning. This step provides interpretability: the model's decision can be traced to specific prototype examples. The final prediction is obtained via $\arg\max \hat{y}$, but the process also includes an adaptive retention mechanism. If the prediction matches the ground truth, the system adds a new prototype to $\mathcal{P}$. This *revise-and-retain* step continuously refines the prototype bank with verified instances, enhancing robustness under distributional shifts.

Overall, Algorithm 2 ensures three desirable properties: (i) **cognitive plausibility**, by reasoning through exemplar similarity; (ii) **interpretability**, as predictions are linked to retrieved cases; and (iii) **adaptivity**, since the prototype bank evolves over time. These properties make the inference process more faithful to human categorization behavior while maintaining practical efficiency.

Algorithm 2 Prototype-Driven MBTI Inference

Require: Test set $U^* = \{post^*\}$; augmentation operator A_2; inference encoder $\mathcal{E}^*$; prototype bank $\mathcal{P}$; similarity operator S; retrieval operator R_k; LLM-based inference operator $\mathcal{I}$
Ensure: Predictions $\{\hat{y}\}$; updated prototype bank $\mathcal{P}$

for all $post^* \in U^*$ **do**
  Apply augmentation: $a' \leftarrow A_2(post^*)$
  Encode representation: $r' \leftarrow \mathcal{E}^*(a')$
  Retrieve prototypes: $\{p_i = \langle a_i, r_i, c_i \rangle\}_{i=1}^{k} \leftarrow R_k(r', \mathcal{P})$ using S
  Infer prediction distribution: $\hat{y} \leftarrow \mathcal{I}(a', \{a_i, c_i\}_{i=1}^{k})$
  Obtain predicted type: $\hat{c} \leftarrow \arg\max \hat{y}$
  if $\hat{c}$ matches ground-truth c **then**
    Construct new prototype: $p' \leftarrow \langle a', r', c \rangle$
    Update bank: $\mathcal{P} \leftarrow \mathcal{P} \cup \{p'\}$
  end if
end for
return predictions $\{\hat{y}\}$ and updated $\mathcal{P}$

Algorithm 3 Prototype Construction

Require: Augmented dataset U_{proto}; compact ($\leq$2B) encoder f_θ (LoRA-enabled)
Ensure: Prototype bank $\mathcal{P}$

Initialize f_θ with LoRA adapters
Train f_θ on U_{proto} with 16-type supervision (details omitted)
$\mathcal{P} \leftarrow \emptyset$
for all $\langle a, c \rangle \in U_{\text{proto}}$ **do**
  Compute embedding $r \leftarrow f_\theta(a)$
  Predict overall MBTI type $\hat{c}$ (for monitoring only)
  Create prototype triplet $p \leftarrow \langle a, r, c \rangle$
  Insert p into prototype bank: $\mathcal{P} \leftarrow \mathcal{P} \cup \{p\}$
end for
Optional: organize $\mathcal{P}$ by class; (e.g., index or shard by c)
return $\mathcal{P}$

G Data Augmentation and Split

Analysis of Table 17. The prompt templates in Table 17 define style-specific instructions for each of the 16 MBTI types. The design rationale is to enable LLMs to generate augmented samples that not only preserve semantic content but also reflect personality-consistent linguistic patterns. Each template encodes a concise description of the target type's stylistic traits (e.g., emotional expressiveness for INFP, logical precision for ISTJ, or energetic spontaneity for ENFP) and provides explicit rewriting instructions. This ensures that generated texts maintain category fidelity while diversifying surface realizations.

Compared to generic augmentation, these prompts introduce *cognitively aligned variation*, grounding synthetic data in psychological theory rather than arbitrary transformations. The resulting augmented corpus therefore exhibits (i) **stylistic fidelity**, where rewritten samples better capture MBTI-consistent tone and expression; (ii) **semantic stability**, since prompts emphasize preservation of meaning while altering style; and (iii) **inter-class contrast**, as differences between MBTI types are explicitly reinforced through tailored instructions. Together, these properties improve the robustness of prototype construction and enhance the interpretability of downstream inference.

Analysis of Table 4. The explanation prompt template in Table 4 is designed to elicit structured, multi-view interpretations of social media posts from an LLM. By framing the model as a psycholinguistics expert, the prompt encourages analysis along three complementary axes: semantic content, sentiment polarity, and linguistic style. The explicit JSON output format enforces consistency, facilitating automatic parsing and integration into downstream pipelines without post-hoc cleaning. This structured approach ensures (i) **semantic grounding**, by summarizing communicative intent; (ii) **affective coverage**, by capturing emotional tone; and (iii) **stylistic profiling**, by characterizing writing mannerisms. Together, these outputs provide rich annotations that enhance prototype construction, improve interpretability of MBTI inference, and enable reproducible evaluation of model behavior across diverse posts.

Role	You are a psycholinguistics expert.
Task	Analyze the following social media post from three perspectives: 1) **Semantic Summary**: main idea or intention. 2) **Sentiment Analysis**: emotions/attitudes. 3) **Linguistic Style**: writing style (e.g., emotional, rational, informal, formal, vague).

Output Format	Return **ONLY** valid JSON with the exact keys below and no extra text:
	{
	"semantic_view": "...",
	"sentiment_view": "...",
	"linguistic_view": "..."
	}

Input Post	<POST_TEXT>

Table 4: LLM explanation prompt template for analyzing social media posts.

Role	You are an expert in MBTI personality typing and linguistic style analysis.
Input	**User Post**: <USER_POST> **Reference Examples**: [Reference Example i] Post Content: <post_casebank> MBTI Type: <type>
Instructions	1. Final Type: _____ 2. Analyze the writing style, tone, logicality, and emotionality. 3. Compare it with each reference example and explain similarities. 4. Conclude with the most likely MBTI type.

Table 5: LLM-based inference prompt template for prototype-driven MBTI classification.

Analysis of Table 5. The inference prompt template specifies how the LLM performs prototype-driven MBTI classification. By positioning the model as an "expert in MBTI personality typing and linguistic style analysis," the template aligns the reasoning process with human expert judgment. The structure explicitly combines the target *user post* with a set of retrieved *reference examples* from the prototype bank, enabling case-based reasoning through direct comparison. The stepwise instructions ensure that predictions are not only label-oriented but also accompanied by linguistic analysis (style, tone, logicality, emotionality) and similarity assessment against exemplars. This design yields three advantages: (i) **faithfulness**, since predictions are grounded in concrete prototype evidence; (ii) **interpretability**, as reasoning steps are made explicit; and (iii) **adaptivity**, because the template can naturally incorporate varying numbers of retrieved cases. Overall, this prompt operationalizes the "retrieve–reason–revise–retain" cycle and provides a cognitively aligned interface between prototype retrieval and LLM inference.

Analysis of Table 6. Table 6 reports the raw distribution of MBTI categories in the Kaggle and Pandora datasets prior to augmentation. Both datasets are highly imbalanced, with a few intuitive patterns. First, introverted intuitive types dominate: INFP and INFJ together account for nearly 38% of Kaggle and 24% of Pandora, while INTP and INTJ cover another 28% and 46% respectively. Conversely, sensing–judging and extroverted sensing types (e.g., ESFJ, ESTJ, ESFP, ESTP) are severely under-represented, each contributing below 1%–2% of total samples. Such skew mirrors broader trends observed in online MBTI communities, where intuitive and introspective users are more active in text-based self-expression.

The imbalance poses two key challenges: (i) **training bias**, as models trained on these datasets may overfit dominant types and underperform on minority ones; and (ii) **generalization risk**, since low-resource classes lack stylistic variety needed for robust prototype construction. These observations motivate the augmentation strategy introduced in Algorithm 1, which aims to achieve class-balanced coverage and stylistic diversity before prototype learning.

MBTI	Kaggle		Pandora	
	Count	Percent	Count	Percent
INTP	1304	15.03%	2336	25.76%
INTJ	1091	12.58%	1847	20.37%
INFP	1832	21.12%	1074	11.85%
INFJ	1470	16.95%	1051	11.59%
ENTP	685	7.90%	631	6.96%
ENFP	675	7.78%	617	6.80%
ISTP	337	3.88%	407	4.49%
ENTJ	231	2.66%	320	3.53%
ISTJ	205	2.36%	195	2.15%
ENFJ	190	2.19%	163	1.80%
ISFP	271	3.12%	123	1.36%
ISFJ	166	1.91%	109	1.20%
ESTP	89	1.03%	72	0.79%
ESFP	48	0.55%	50	0.55%
ESTJ	39	0.45%	43	0.47%
ESFJ	42	0.48%	29	0.32%
Total	**8675**	**100%**	**9067**	**100%**

Table 6: Distribution of MBTI types in **Kaggle** and **Pandora** datasets *before* augmentation

MBTI	Kaggle (Aug)		Pandora (Aug)	
	Count	Percent	Count	Percent
INTP	2144	6.30%	2336	6.87%
INTJ	2115	6.21%	2147	6.32%
INFP	2235	6.56%	2155	6.34%
INFJ	2120	6.23%	2142	6.30%
ENTP	2120	6.23%	2127	6.26%
ENFP	2117	6.22%	2111	6.21%
ISTP	2102	6.17%	2106	6.20%
ENTJ	2105	6.18%	2088	6.14%
ISTJ	2062	6.06%	2093	6.16%
ENFJ	2126	6.24%	2104	6.19%
ISFP	2188	6.43%	2098	6.17%
ISFJ	2103	6.18%	2116	6.23%
ESTP	2148	6.31%	2102	6.19%
ESFP	2068	6.07%	2077	6.11%
ESTJ	2120	6.23%	2090	6.15%
ESFJ	2177	6.39%	2102	6.19%
Total	**34050**	**100%**	**33994**	**100%**

Table 7: Distribution of MBTI types in **Kaggle** and **Pandora** datasets *after* augmentation

Analysis of Table 7. Table 7 presents the MBTI distributions in Kaggle and Pandora after applying the proposed augmentation procedure. In contrast to the skewed pre-augmentation distributions (Table 6), the post-augmentation datasets exhibit near-uniform coverage across all 16 types. Each type constitutes approximately 6% of the total, with only minor fluctuations (within $\pm 0.3\%$).

This balanced distribution addresses the two major issues observed earlier: (i) **class imbalance** is mitigated, ensuring that minority types such as ESFJ, ESTJ, and ESFP are equally represented alongside dominant types like INFP and INTJ; and (ii) **stylistic diversity** is enhanced by multi-dimensional augmentation, which increases variability within each class without inflating dataset size arbitrarily.

As a result, the augmented datasets provide a more equitable training signal for prototype construction, reducing the risk of bias toward majority classes and improving cross-class generalization. This uniformity also simplifies downstream evaluation by aligning per-type accuracy with overall performance, making improvements interpretable and comparable across categories.

Analysis of Table 8. Table 8 compares the distributions of the four MBTI dimensions before and after augmentation for Kaggle and Pandora. In the pre-augmentation setting, both datasets exhibit strong biases: introversion (I) dominates over extraversion (E) with ratios exceeding 3:1; intuition (N) heavily outweighs sensing (S), particularly in Pandora where nearly 89% of users fall into the N pole; thinking (T) and feeling (F) distributions are skewed differently across datasets, with T-dominance in Pandora and F-dominance in Kaggle; and perceiving (P) is systematically overrepresented compared to judging (J). These imbalances reflect community-level self-selection effects, as certain personality types are more active in online MBTI forums.

After augmentation, each pole within a dimension is balanced close to 50%–50%, with deviations under 0.5%. This equilibrium ensures that the augmented datasets no longer privilege one side of a dichotomy, thereby reducing systemic bias in downstream classifiers. Importantly, balancing at the dimension level complements type-level augmentation (Table 7): while type-level balancing equalizes the 16 categories, dimension-level balancing guarantees consistent representation of the four psychological dichotomies. Together, these adjustments provide a more cognitively plausible and statistically robust foundation for prototype construction and MBTI inference.

Analysis of Table 9. Table 9 summarizes the dataset splits for Kaggle and Pandora after augmentation. The design follows two principles. First, the

Dimension	Pole	Kaggle		Pandora	
		Count (Pre / Aug)	Percent	Count (Pre / Aug)	Percent
E / I	E	1999 / 16963	23.04% / 49.8%	1925 / 16870	21.23% / 49.6%
	I	6676 / 17069	76.96% / 50.2%	7142 / 17124	78.77% / 50.4%
S / N	S	1197 / 16968	13.80% / 49.9%	1028 / 16902	11.34% / 49.7%
	N	7478 / 17064	86.20% / 50.1%	8039 / 17112	88.66% / 50.3%
T / F	T	3981 / 16898	45.89% / 49.7%	5851 / 17021	64.53% / 50.0%
	F	4694 / 17134	54.11% / 50.3%	3216 / 16993	35.47% / 50.0%
J / P	J	3434 / 16928	39.59% / 49.7%	3757 / 16972	41.44% / 49.9%
	P	5241 / 17104	60.41% / 50.3%	5310 / 17042	58.56% / 50.1%
Total		8675 / 34068	100% / 100%	9067 / 34079	100% / 100%

Table 8: Distribution over the four MBTI dimensions in Kaggle and Pandora datasets before and after augmentation

train and *validation* sets are constructed from the augmented corpora to ensure class balance across all four MBTI dichotomies. This prevents learning bias toward majority poles and provides stable supervision signals during prototype construction. Second, the *test* sets remain unaugmented, preserving the natural class skew observed in the raw data. This choice makes evaluation more realistic, as models must generalize to authentic distributions rather than artificially balanced ones.

Across both datasets, training and validation counts are tightly matched across poles (differences $< 1\%$), reflecting the success of augmentation in balancing the data. In contrast, the test sets reveal the original imbalances (e.g., far more N than S, and more I than E), which allows us to assess robustness under distribution shift. This split strategy thus provides (i) **fair training**, with balanced supervision signals; (ii) **realistic evaluation**, by retaining natural skew in the test data; and (iii) **generalization stress-testing**, by forcing models trained on balanced data to handle unbalanced distributions at inference time.

H Baselines

We select several representative baseline methods in our experiments, ranging from traditional machine learning approaches to deep learning architectures and the latest large language model (LLM)-based methods. These baselines not only reflect the developmental trajectory of personality detection research but also provide a solid comparative foundation for evaluating ProtoMBTI.

Traditional machine learning methods. SVM (Cui and Qi, 2017) and XGBoost (Tadesse et al., 2018) are widely used in early personality detection studies. These methods typically concatenate all user posts into a single long document, extract statistical features using a bag-of-words model, and then apply classification algorithms such as SVM or XGBoost for prediction. The advantages of these methods lie in their simplicity and low computational cost, but they fail to capture semantic information and contextual relationships effectively.

Neural network methods. BiLSTM (Tandera et al., 2017) model the contextual information of text by employing bidirectional LSTM networks, and they merge post embeddings into a unified representation using average pooling for personality prediction. Compared with traditional methods, BiLSTM provides stronger sequence modeling capability, yet it still struggles with long-text modeling and global semantic understanding.

Pretrained language models such as BERT_mean (Keh et al., 2019) and BERT_concat (Jiang et al., 2020) introduce transformer-based architectures into personality detection tasks. BERT_mean encodes each post with BERT and applies average pooling to generate a user-level representation, which is then mapped to personality labels. BERT_concat concatenates all user posts into a single long document, encodes the text with BERT, and then applies fully connected layers for classification. Both approaches significantly improve semantic modeling capacity, but they remain limited in capturing personality consistency across multiple posts.

AttRCNN (Xue et al., 2018) employs a hierarchical deep neural network that combines an AttRCNN structure with an Inception variant to capture deep semantic features, while also incorporating statistical linguistic features to enhance recognition accuracy. AttnSeq (Lynn et al., 2020) introduces a hierarchical attention mechanism that applies at-

Dataset	Types	Train(Aug)	Validation(Aug)	Test(Raw)
Kaggle	I / E	13656 / 13568	1706 / 1698	652 / 201
	S / N	13650 / 13574	1697 / 1707	111 / 742
	T / F	13520 / 13704	1689 / 1705	403 / 450
	P / J	13682 / 13542	1711 / 1693	514 / 339
Pandora	I / E	13820 / 13470	1720 / 1690	480 / 355
	S / N	13710 / 13580	1705 / 1690	118 / 725
	T / F	13560 / 13730	1690 / 1705	395 / 460
	P / J	13680 / 13610	1708 / 1692	505 / 340

Table 9: Dataset Splits for Kaggle and Pandora Datasets (After Augmentation)

tention at both the word level and the message level, enabling the model to capture personality-related signals at multiple granularities. These approaches partly alleviate the challenges of long-text modeling and emphasize the contributions of different semantic levels. Transformer-MD (Yang et al., 2021a) is specifically designed for multi-document personality detection. It employs a Multi-Document Transformer architecture with memory tokens and shared positional embeddings, allowing dynamic information access across posts, mitigating order bias, and constructing coherent personality representations over multiple documents.

TrigNet (Yang et al., 2021b) integrates psycholinguistic knowledge by introducing a psycholinguistic tripartite graph network. This method combines a BERT-based initializer with a graph attention mechanism to incorporate psycholinguistic features into the task, significantly enhancing the model's ability to capture the relationship between language use and personality traits.

D-DGCN (Yang et al., 2023) further proposes a Dynamic Deep Graph Convolutional Network that models user posts as dynamic graphs with posts as nodes. It captures cross-post relationships through multi-hop connectivity and deep graph convolution layers. This approach reduces the influence of post order bias and improves the robustness of personality feature representations.

LLM-based methods. TAE (Hu et al., 2024) applies large language models for data augmentation and combines them with smaller models for efficient inference. Its central idea is to leverage the generative capability of LLMs in semantic, affective, and linguistic dimensions to augment posts, thereby improving downstream training effectiveness and generalization. ETM (Bi et al., 2025) further exploits the dual capability of LLMs in personality detection, using them both as generators for synthesizing high-quality training samples and as embedding extractors for semantically rich representations. This approach enhances performance in data-scarce scenarios and demonstrates the potential of LLMs in this domain.

Qwen vanilla and in-context learning. Table 10 summarizes the prompt templates used for the Qwen-based LLM baselines under both vanilla and in-context learning (ICL) settings. The vanilla prompt directly queries MBTI dimension predictions, while the ICL prompt conditions the model on exemplar-driven language styles corresponding to each MBTI type. All prompts are kept fixed across datasets to ensure a fair and controlled comparison.

Source of MBTI knowledge. The language-style descriptions associated with each MBTI type are grounded in officially published MBTI theory and practitioner-oriented materials released by the Myers–Briggs Foundation. They reflect canonical interpretations of preference dimensions and are reformulated to characterize language style and reasoning patterns rather than explicit personality labels.

Source of examples. All example sentences are manually constructed illustrative instances for experimental prompting purposes. They are not direct quotations from official MBTI materials and are intended solely to demonstrate typical language styles in a neutral and non-stereotypical manner.

ISTJ. Individuals with this preference are typically characterized by a concise, factual, and structured language style, emphasizing responsibility, established procedures, and reliability, with minimal emotional expression. An illustrative example is: "Based on the available information and established procedures, this option appears to be the most reliable."

ISFJ. This language style is warm yet reserved, focusing on practical support and attentiveness to others' needs, conveyed through a polite and con-

Setting	Prompt Template

Vanilla	`You are a professional MBTI personality type analyst. Please analyze the author's MBTI personality type across four dimensions based on the following text content: E/I (Extroversion/Introversion), S/N (Sensing/Intuition), T/F (Thinking/Feeling), J/P (Judging/Perceiving).`
	`Text: {text}`
	`Please output the results in the following format. Only output the letters for the four dimensions, with no other content: E/I: X S/N: X T/F: X J/P: X`

ICL	`You are given examples of language produced by individuals with different MBTI personality types. MBTI types reflect stable preferences in: - Energy orientation (Extroversion vs. Introversion) - Information processing (Sensing vs. Intuition) - Decision-making (Thinking vs. Feeling) - Lifestyle orientation (Judging vs. Perceiving)` `IMPORTANT CONSTRAINTS: 1. You must NOT explicitly mention MBTI, personality labels, or typology terms in your response. 2. You must express personality ONLY through language style, tone, structure, and reasoning patterns. 3. Focus on HOW the person speaks, not on describing the personality. 4. Stay consistent with the speaking style demonstrated in the examples. 5. Do not exaggerate or caricature the style. Below are reference examples illustrating the typical speaking styles of each MBTI type.` `[Insert the 16-type style exemplars here, unchanged.]` `TASK: Now respond to the user query as if you were a person with MBTI type: <MBTI_TYPE>. Follow the corresponding language style strictly.`

Table 10: Qwen Baseline Prompts under Vanilla and In-Context Learning (ICL) Settings.

siderate tone. An illustrative example is: "This approach should help everyone feel more comfortable and reduce unnecessary pressure."

INFJ. The language style is reflective and abstract, with an emphasis on meaning, long-term impact, and underlying intent, expressed in a calm and thoughtful manner. An illustrative example is: "What matters most is whether this decision aligns with the direction we ultimately want to move toward."

INTJ. This style is strategic and analytical, future-oriented, and focused on systems and long-term planning, while remaining emotionally restrained. An illustrative example is: "If the objective is sustainable progress, this step follows logically."

ISTP. The language style is direct and pragmatic, oriented toward problem-solving and immediate functionality, often with minimal explanation. An illustrative example is: "Let's test this first and adjust if necessary."

ISFP. This style is gentle and personal, valuing authenticity and individual preference, and generally avoiding confrontation. An illustrative example is: "I feel more at ease with this option; it seems right to me."

INFP. The language style is value-driven and idealistic, expressing internal alignment and personal meaning with emotionally sincere but understated language. An illustrative example is: "If this choice conflicts with what I believe in, it would be difficult for me to support it."

INTP. This language style is abstract and theoretical, relying on hypotheses and conceptual reasoning while maintaining emotional neutrality. An illustrative example is: "From a theoretical perspective, several interpretations are possible."

ESTP. The style is action-oriented and decisive, focused on immediacy and real-world impact, often conveyed with an energetic tone. An illustrative example is: "Let's move forward and see what happens instead of overanalyzing."

ESFP. This language style is expressive and present-focused, emphasizing experience, engagement, and openness in emotional expression. An illustrative example is: "This would make the experience more engaging for everyone involved."

ENFP. The style is enthusiastic and possibility-focused, using imaginative and future-oriented language with a positive and inspiring tone. An illustrative example is: "This could open up several new directions we haven't explored yet."

ENTP. This language style is playful and debate-oriented, characterized by questioning assumptions and reframing perspectives. An illustrative example is: "What if we approached this from a completely different angle?"

ESTJ. The language style is assertive and struc-

tured, emphasizing efficiency, rules, and execution, conveyed with a confident tone. An illustrative example is: "This is the most efficient approach and should be implemented accordingly."

ESFJ. This style is cooperative and socially attentive, focusing on harmony, consensus, and shared understanding, with a warm yet organized tone. An illustrative example is: "We should choose an option that everyone feels comfortable supporting."

ENFJ. The language style is encouraging and people-focused, emphasizing growth, motivation, and collective purpose through emotionally supportive language. An illustrative example is: "I believe this decision can help everyone move forward more confidently."

ENTJ. This style is decisive and goal-oriented, focused on vision, leadership, and outcomes, expressed in a direct and authoritative manner. An illustrative example is: "The goal is clear; execution is what matters now."

I More Results and Analysis

Performance Comparison Our proposed ProtoMBTI framework surpasses all existing methods across all metrics, as shown in Table 11, and achieves the best generalization performance on mixed datasets. *Comparison on single datasets.* For the four MBTI dimensions, ProtoMBTI$_{\text{Qwen}}$ achieves an average accuracy of 85.14% on the Kaggle dataset, significantly higher than the previous best model ETM (77.79%). On the Pandora dataset, ProtoMBTI$_{\text{GPT4o}}$ reaches 71.41%, again exceeding ETM (65.77%). For the 16-type classification task, the best theoretical value reported in prior work is only 35.89% (ETM) on Kaggle, while ProtoMBTI$_{\text{Qwen}}$ achieves 71.42%, representing a remarkable improvement. These results demonstrate that under single-dataset settings, ProtoMBTI outperforms the current state-of-the-art methods in both four-dimension and 16-type classification. For LLM-based baselines, vanilla prompting consistently outperforms ICL across both datasets and all four dichotomies, suggesting that the subtle, fuzzy, and implicit nature of MBTI personality traits may not be well suited to in-context learning.

Comparison on mixed datasets. When Kaggle and Pandora are combined for training while validation and test sets remain consistent within a single dataset, ProtoMBTI$_{\text{mix}}$ performs substantially better than single-dataset training. For the 16-type accuracy, performance on Kaggle increases from 71.41% to 85.54%, and on Pandora from 60.22% to 92.13%. The model also achieves the best performance across all four MBTI dimensions under this setting.

Cross-dataset evaluation. When validation and test sets are swapped across datasets, ProtoMBTI$_{\text{mix-ex}}$ still maintains strong generalization. For the 16-type classification, the model achieves 81.23% on Kaggle and 81.15% on Pandora. For the four-dimension average, Kaggle reaches 91.62%, only 1.88 percentage points lower than the same-domain ProtoMBTI$_{\text{mix}}$, and Pandora reaches 90.87%, 5.54 percentage points lower. Although performance decreases compared to the same-domain setting, it remains far superior to single-dataset training.

Table 2 presents the ablation results on the Kaggle dataset, reporting accuracies for the four MBTI dimensions, the average accuracy, and the 16-type classification accuracy. The table is organized into three parts: the full ProtoMBTI model, ablations on the prototype bank and data augmentation strategies, and encoder-only baselines without prototype reasoning.

Effect of prototype selection. The results show that appropriate prototype selection is critical to model performance. ProtoMBTI$_{\text{[k+1,2k]}}$ achieves an average accuracy of 80.69%, slightly lower than the full ProtoMBTI$_{\text{Qwen}}$ (85.14%), indicating that using secondary prototypes results in limited degradation. In contrast, ProtoMBTI$_{\text{RandomProto}}$ and ProtoMBTI$_{\text{Semantic}}$ reduce the average accuracy to 79.66% and 79.50%, respectively, and further decrease the 16-type accuracy to 50.77% and 50.15%. Notably, both perform worse than ProtoMBTI$_{\text{ZeroProto}}$, which excludes prototypes entirely, suggesting that inappropriate prototype selection can interfere with classification.

Effect of data augmentation. Removing data augmentation components leads to a substantial performance drop. ProtoMBTI$_{\text{Explain_only}}$ achieves a 16-type accuracy of 56.62%, while ProtoMBTI$_{\text{Raw}}$ further decreases to an average accuracy of 77.96% and a 16-type accuracy of 45.32%. These results indicate that category balancing and explanation-based augmentation are both important, particularly for fine-grained 16-type classification.

Effect of prototype reasoning. Performance degrades markedly when prototype reasoning is removed and classification relies solely on encoder

Methods	Kaggle						Pandora					
	I/E	S/N	T/F	P/J	Avg.	16-Type	I/E	S/N	T/F	P/J	Avg.	16-Type
SVM	53.34	47.75	76.72	63.03	60.21	12.32	44.74	46.92	64.62	56.32	53.15	7.64
XGBoost	56.67	52.85	75.42	65.94	62.72	14.89	45.99	48.93	63.51	55.55	53.50	7.94
BiLSTM	57.82	57.87	69.97	57.01	60.67	13.35	48.01	52.01	63.48	56.21	54.93	8.91
$\text{BERT}_{\text{concat}}$	58.33	53.88	69.36	60.88	60.61	13.27	54.22	49.15	58.31	53.14	53.71	8.26
$\text{BERT}_{\text{mean}}$	64.65	57.12	77.95	65.25	66.24	18.78	56.60	48.71	64.70	56.07	56.52	10.00
AttRCNN	59.74	64.08	78.77	66.44	67.25	20.03	48.55	56.19	64.39	57.26	56.60	10.06
SN+Attn	65.43	62.15	78.05	63.92	67.39	20.29	56.98	54.78	60.95	54.81	56.88	10.43
Transformer-MD	66.08	69.10	79.19	67.50	70.47	24.41	55.26	58.77	69.26	60.90	61.05	13.70
TrigNet	69.54	67.17	79.06	67.69	70.86	25.00	56.69	55.57	66.38	57.27	58.98	11.98
D-DGCN	68.41	65.66	79.56	67.22	70.21	24.02	61.55	55.46	71.07	59.96	62.01	14.55
D-DGCN+ℓ_0	69.52	67.19	80.53	68.16	71.35	25.64	59.98	55.52	70.53	59.56	61.40	13.99
GPT3.5	65.86	51.69	78.60	63.93	66.89	17.11	55.52	49.79	71.25	60.51	59.27	11.92
$\text{Vanilla}_{\text{Qwen}}$	58.86	56.17	58.41	53.53	56.74	10.30	51.47	52.01	56.84	55.05	53.84	8.39
ICL_{Qwen}	50.23	44.31	47.86	48.19	47.65	5.14	46.18	47.57	53.30	52.26	49.83	6.13
TAE	70.90	66.21	81.17	70.20	72.07	26.75	62.57	61.01	70.53	59.34	63.05	15.98
ETM	68.97	71.21	86.19	84.78	77.79	35.89	68.57	64.91	66.07	63.53	65.77	18.68
$\text{ProtoMBTI}_{\text{llama}}$	81.92	87.70	86.04	82.47	84.03	71.11	69.05	68.85	68.98	70.82	69.43	50.30
$\text{ProtoMBTI}_{\text{Qwen}}$	83.74	88.10	84.54	84.18	85.14	71.42	71.63	66.98	73.25	70.33	70.55	41.86
$\text{ProtoMBTI}_{\text{GPT4o}}$	82.36	85.55	82.70	80.04	82.66	68.39	70.41	70.65	73.32	71.27	71.41	60.22
$\text{ProtoMBTI}_{\text{mix}}$	**93.54**	**93.54**	**95.69**	**93.23**	**93.50**	**85.54**	**96.25**	**97.08**	**96.75**	**95.54**	**96.41**	**92.13**
$\text{ProtoMBTI}_{\text{mix-ex}}$	91.08	90.77	94.46	90.15	91.62	81.23	90.44	91.25	91.35	90.44	90.87	81.15

Table 11: Performance comparison of ProtoMBTI and baselines on Kaggle and Pandora datasets. Metrics include four dimension accuracies, their average, and the 16-type accuracy (theoretical for baselines computed as the product of the four dimension accuracies, direct prediction for ProtoMBTI). Subscripts denote different LLMs, *mix* for same-source training/testing, and *mix-ex* for cross-source evaluation.

representations. EncoderOnly models achieve at most 60.62% accuracy on the 16-type task, which is substantially lower than ProtoMBTI models with prototype reasoning (up to 71.42%). This suggests that direct embedding-based classification is insufficient to capture the categorical structure of MBTI.

Sensitivity of the 16-type metric. The 16-type classification accuracy is more sensitive to ablation than the four-dimension average accuracy. While $\text{ProtoMBTI}_{\text{Qwen}}$ reaches 71.42% on the 16-type task, $\text{ProtoMBTI}_{\text{Raw}}$ drops to 45.32%, resulting in a gap of 26.1 percentage points. In comparison, variations in the four-dimension average accuracy are relatively smaller, indicating that fine-grained classification amplifies the effects of model components.

Analysis of Table 12. Table 12 reports the performance of different backbone encoders within the proposed 4D Classifier framework on the Kaggle validation set. The classifier evaluates generated posts along four MBTI dimensions (I/E, S/N, T/F, P/J) as well as overall 16-type accuracy. Results show that all three transformer-based variants (BERT, RoBERTa, DeBERTa) achieve strong dimensional classification, with accuracies exceeding 83% across all dichotomies. Among them, 4D-DeBERTa yields the best overall performance, reaching 88.63% average dichotomy accuracy and 71.08% 16-type accuracy.

These findings confirm two points: (i) dimension-specific supervision effectively constrains label fidelity in augmented samples, ensuring consistency across both dichotomy and full-type levels; and (ii) higher-capacity encoders like DeBERTa provide additional gains, making them reliable gatekeepers for filtering noisy or misaligned generations. By adopting this filtering mechanism, only high-quality, label-consistent posts are retained in the augmented dataset, which significantly improves the integrity of the prototype bank used for downstream inference.

Methods	Kaggle					
	I/E	S/N	T/F	P/J	Avg	16-type
BERT	88.59	92.05	84.22	83.87	87.18	67.28
Roberta	89.17	92.04	85.48	86.52	88.30	69.93
Deberta	89.63	93.09	85.60	86.18	88.63	71.08

Table 12: Performance comparison on Kaggle validation set.

Analysis of Table 13. Table 13 compares two LLM variants, 4o and 4o-mini, on post-level augmentation quality across MBTI types. The "ratio" row shows the number of generated posts that suc-

cessfully passed the 4D classifier filtering, while "Acc. score" reflects the average acceptance rate across types. Results indicate that 4o-mini consistently outperforms 4o, achieving a higher overall pass ratio (108 vs. 99) and a +0.0584 improvement in acceptance score.

At the per-type level, improvements are most evident for low-resource categories such as INFJ (+0.20) and INFP (+0.30), where stylistic fidelity is harder to capture. Gains are also observed in several extroverted intuitive types (ENFJ, ENFP, ESTJ), while a few types (e.g., ESFP, ISFP, ISTP) exhibit small drops. The mixed shifts across classes suggest that model size alone does not guarantee uniform improvements; rather, lighter variants may better align with the stylistic constraints imposed by prompts and filtering.

Overall, these results demonstrate that 4o-mini provides a more effective balance between generation diversity and label consistency, making it a preferable choice for large-scale augmentation in our framework.

MBTI	**4o**	**4o-mini**	**Δ**
ratio	99 / 154	108 / 154	+9
Acc. score	0.6429	0.7013	+0.0584
ENFJ	0.40	0.60	+0.20
ENFP	0.70	0.90	+0.20
ENTJ	0.90	1.00	+0.10
ENTP	0.80	0.80	0.00
ESFJ	1.00	1.00	0.00
ESFP	0.90	0.70	-0.20
ESTJ	0.70	0.90	+0.20
ESTP	0.70	0.80	+0.10
INFJ	0.10	0.30	+0.20
INFP	0.10	0.40	+0.30
INTJ	0.30	0.30	0.00
INTP	0.50	0.50	0.00
ISFJ	0.80	0.80	0.00
ISFP	0.70	0.60	-0.10
ISTJ	0.90	0.90	0.00
ISTP	1.00	0.90	-0.10

Table 13: Performance comparison of different LLMs (4o vs 4.1mini) in post-level data augmentation across MBTI types. Δ indicates the performance gap (4omini - 4o).

Analysis of Figure 6. Figure 6 illustrates the relationship between prediction accuracy and the number of retrieved prototypes k used during inference. The results show that accuracy improves substantially when increasing k from 1 to 3, reaching the highest performance at $k = 3$ (59.38%). Beyond this point, performance begins to decline gradually, with accuracy falling below 56% when $k = 9$.

This trend highlights a key trade-off in prototype aggregation. Using too few prototypes (e.g., $k = 1$) provides insufficient context and may lead to unstable predictions dominated by a single exemplar. Conversely, using too many prototypes (e.g., $k = 9$) introduces noise and dilutes the discriminative signal, as irrelevant or weakly similar cases are included. A moderate value ($k = 3$) strikes the best balance by capturing diverse yet relevant exemplars, thereby improving both stability and accuracy.

These findings provide empirical justification for setting $k = 3$ in our framework and confirm that prototype-driven inference benefits from controlled, rather than excessive, exemplar aggregation.

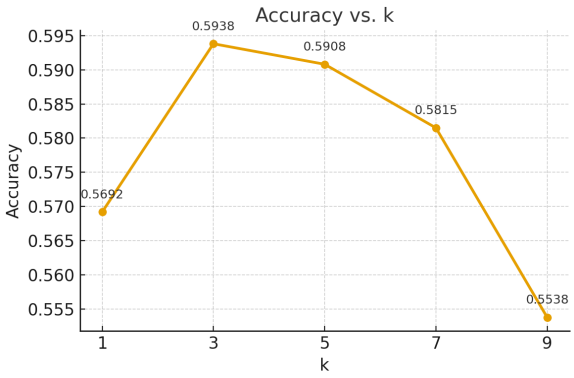

Figure 6: **Accuracy with different values of k in prototype selection.** The model achieves the best performance at k=3 (59.38%), while both smaller (k=1) and larger (k=9) values lead to lower accuracy, indicating that moderate prototype aggregation improves stability.

Analysis of Table 14. While the main text reports results in terms of accuracy, Table 14 provides complementary evaluation using the macro-averaged F1 score, which is more sensitive to class imbalance and therefore a stricter measure of performance. The results reveal several consistent patterns. First, single-backbone variants of ProtoMBTI (LLaMA, Qwen, GPT4o) achieve moderate F1 scores across the four MBTI dimensions (typically 75–85%) but exhibit sharp drops in the 16-type setting (33–63%), reflecting difficulty in handling fine-grained categories under limited per-class support.

In contrast, the ensemble-based variants (ProtoMBTI$_{\text{mix}}$ and ProtoMBTI$_{\text{mix-ex}}$) show substantial improvements. ProtoMBTI$_{\text{mix}}$ achieves the strongest results overall, reaching above 90% F1 across all four dimensions and exceeding 85% on Kaggle and 92% on Pandora for the 16-type task. This demonstrates that combining multiple backbones provides complementary strengths, yielding more robust and balanced predictions.

ProtoMBTI$_{\text{mix-ex}}$, though slightly weaker, still outperforms all single-backbone baselines by a large margin.

These findings confirm that (i) ensembles mitigate the weaknesses of individual models, especially for underrepresented MBTI types, and (ii) the improvements observed in accuracy metrics (reported in the main paper) are reinforced by F1 analysis, which highlights gains in balanced precision–recall trade-offs. Thus, ProtoMBTI's ensemble design not only boosts overall correctness but also ensures fairness and robustness across the MBTI label space.

Methods	I/E	S/N	T/F	P/J	16-Type
ProtoMBTI$_{\text{llama}}$	76.55	76.55	82.26	74.78	54.89
ProtoMBTI$_{\text{Qwen}}$	80.36	83.36	84.54	81.86	62.04
ProtoMBTI$_{\text{GPT4o}}$	79.96	77.25	86.07	77.49	57.94
ProtoMBTI$_{\text{mix}}$	88.55	98.14	95.91	93.48	85.54
ProtoMBTI$_{\text{mix-ex}}$	90.46	92.17	91.08	90.17	90.97

Table 15: Recall (%) of ProtoMBTI variants on the Kaggle dataset, including 16-type classification.

Methods	I/E	S/N	T/F	P/J	16-Type
ProtoMBTI$_{\text{llama}}$	68.16	66.41	72.46	62.11	33.20
ProtoMBTI$_{\text{Qwen}}$	73.31	68.73	73.06	68.03	31.45
ProtoMBTI$_{\text{GPT4o}}$	68.95	64.84	68.95	63.67	31.45
ProtoMBTI$_{\text{mix}}$	96.83	97.12	96.37	95.71	96.51
ProtoMBTI$_{\text{mix-ex}}$	82.44	98.14	94.74	90.58	91.23

Table 16: Recall (%) of ProtoMBTI variants on the Pandora dataset, including 16-type classification.

Analysis of Table 15 and Table 16 Table 15 and Table 16 reports recall scores for all ProtoMBTI variants on Kaggle and Pandora. Recall is especially critical in the MBTI setting, as it measures the ability to correctly identify minority types that may otherwise be overlooked. Consistent with accuracy and F1 trends, single-backbone models (LLaMA, Qwen, GPT4o) achieve reasonable recall on the four MBTI dimensions (typically 75–85%), but their performance drops sharply in the 16-type setting (≈31–62%), indicating that many fine-grained categories are missed.

The ensemble approaches again deliver clear improvements. ProtoMBTI$_{\text{mix}}$ achieves the highest recall overall, surpassing 90% across dimensions and reaching 96.51% on Pandora for the 16-type task. ProtoMBTI$_{\text{mix-ex}}$ also performs strongly, particularly on Kaggle (90.97% in 16-type recall), confirming its robustness. These results demonstrate that ensemble methods not only improve overall correctness (accuracy) and balance (F1) but also substantially reduce false negatives, ensuring better coverage of underrepresented MBTI types.

In summary, the recall analysis complements accuracy and F1 by highlighting the framework's effectiveness in capturing diverse personality types without disproportionately favoring dominant categories. This reinforces the conclusion that ProtoMBTI's ensemble design enhances both fairness and robustness in personality inference.

Overview of Figure 7, Figure 8, and Figure 9. The confusion matrices and ROC analyses show that ProtoMBTI learns stable personality prototypes across models and datasets. In-domain evaluations exhibit clear diagonal dominance, while cross-domain and heterogeneous settings introduce increased confusion, indicating sensitivity to domain shift. Training on the mixed Kaggle–Pandora dataset substantially mitigates these effects, yielding more robust prototype boundaries. ROC results further confirm consistently strong and balanced performance across MBTI categories. Overall, prototype-informed learning enables generalizable MBTI inference under heterogeneous data distributions, though fine-grained distinctions among closely related types remain challenging.

Methods	Kaggle					Pandora				
	I/E	S/N	T/F	P/J	16-Type	I/E	S/N	T/F	P/J	16-Type
ProtoMBTI$_{\text{llama}}$	77.62	78.80	81.82	75.39	56.52	67.26	64.59	72.21	57.54	36.27
ProtoMBTI$_{\text{Qwen}}$	81.30	85.07	84.31	82.50	63.44	71.99	65.91	72.92	67.33	34.22
ProtoMBTI$_{\text{GPT4o}}$	80.61	79.74	85.85	78.26	60.33	68.90	62.87	68.93	60.26	33.93
ProtoMBTI$_{\text{mix}}$	91.70	95.26	95.91	92.14	85.59	96.27	97.08	96.74	95.55	92.11
ProtoMBTI$_{\text{mix-ex}}$	90.44	91.33	91.33	90.41	81.18	88.16	93.36	94.74	88.65	81.48

Table 14: Overall performances of ProtoMBTI variants in F1 (%), including 16-type classification.

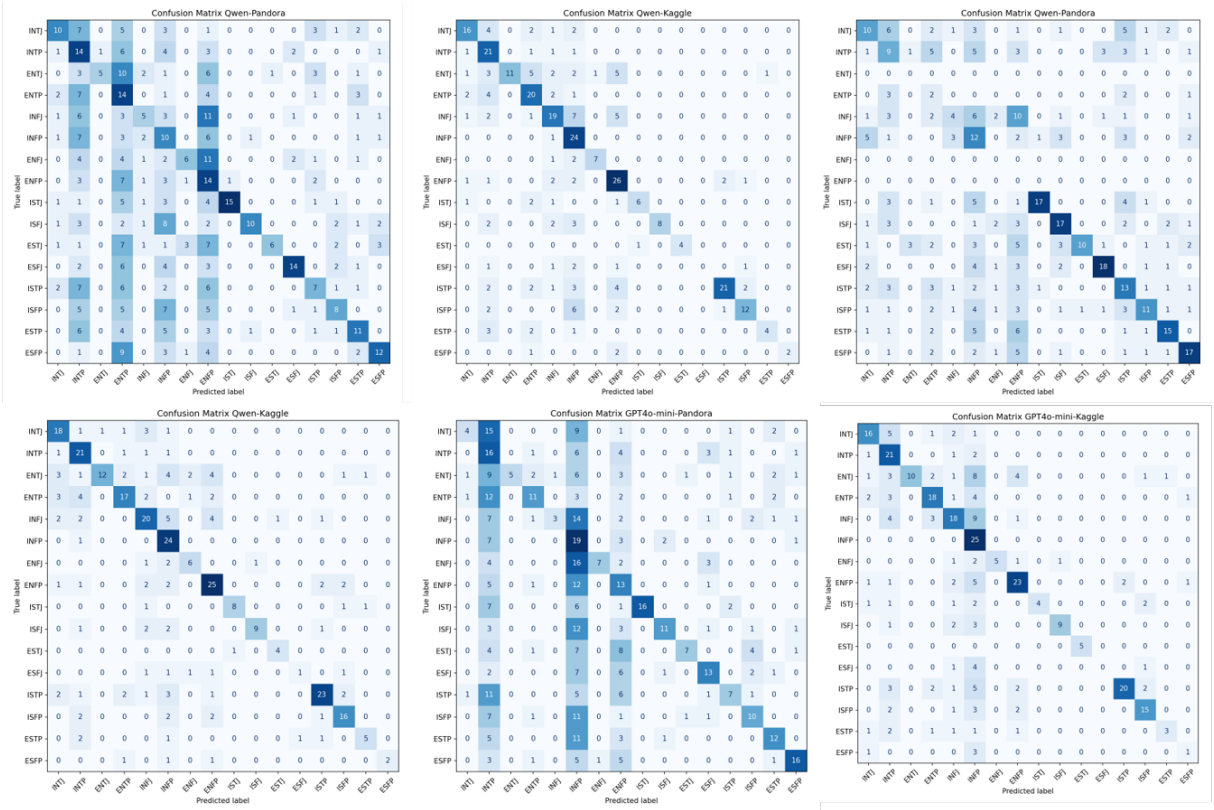

Figure 7: Confusion matrices of ProtoMBTI across different backbone models and datasets. The figure reports 16-class MBTI confusion matrices for ProtoMBTI instantiated with **GPT-4o-mini**, **Qwen**, and **LLaMA** backbones, evaluated on the **Kaggle** and **Pandora** datasets. Across in-domain settings (Kaggle→Kaggle and Pandora→Pandora), all models exhibit clear diagonal dominance for several prototypical personality types such as *INFP*, *ENFP*, and *ISTP*, indicating strong recognition of their characteristic linguistic patterns. However, systematic misclassifications frequently arise between semantically adjacent types, including *ENTP* vs. *ENTJ* and *INFJ* vs. *INTP*, reflecting the intrinsic difficulty of distinguishing categories with overlapping cognitive and discourse traits. Under the more challenging Pandora benchmark and cross-domain scenarios, diagonal dominance is noticeably weakened for all backbones, with substantially increased confusion across nearly all categories. In particular, ProtoMBTI$_{\text{LLaMA}}$ shows pronounced sensitivity to domain shift, where even highly prototypical classes such as *INFP*, *ENFP*, and *ISFP* lose discriminative clarity. Overall, these results indicate that while ProtoMBTI effectively captures core personality prototypes across models, resolving fine-grained distinctions among closely related MBTI types remains challenging, underscoring the necessity of prototype-aware disambiguation and robust transfer mechanisms for stable personality inference under heterogeneous data distributions.

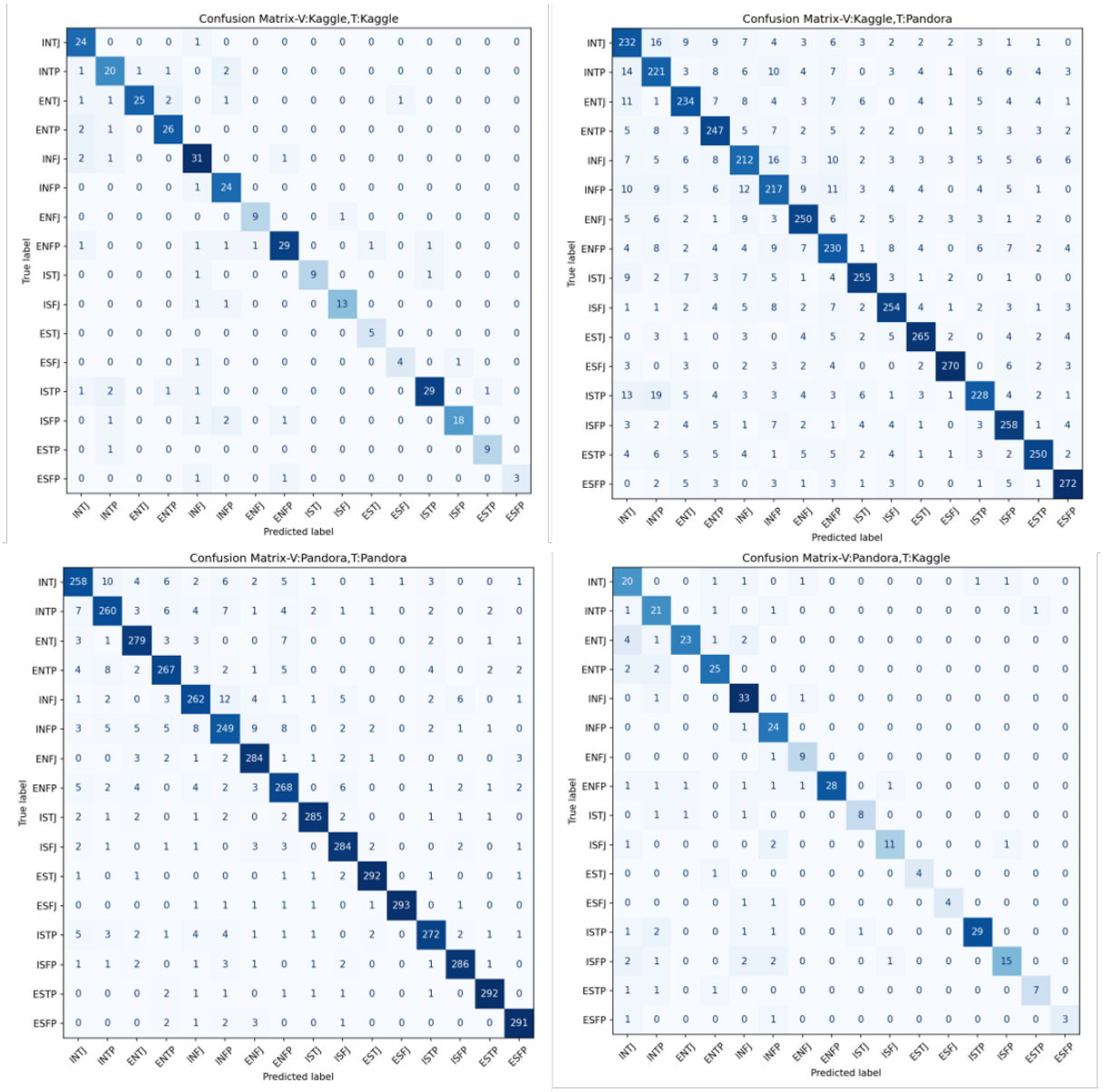

Figure 8: Confusion matrices of ProtoMBTI trained on the mixed Kaggle–Pandora dataset under different validation–test configurations. The figure reports four settings: **Mix–Kaggle–Kaggle**, **Mix–Pandora–Pandora**, **Mix–Kaggle–Pandora**, and **Mix–Pandora–Kaggle**. Across within-domain evaluations (Kaggle→Kaggle and Pandora→Pandora), the model exhibits exceptionally strong diagonal dominance, with several personality types such as *INFJ*, *ENFP*, *ISTP*, *ENFJ*, *ESFJ*, and *ESTP* achieving near-perfect recognition. Misclassifications in these settings are sparse and largely confined to semantically adjacent categories (e.g., *INTJ* vs. *INTP*), indicating that mixed-domain training enables ProtoMBTI to learn well-separated and robust prototypical boundaries. Under cross-domain evaluations, diagonal dominance is slightly weakened and confusions increase, particularly among closely related types such as *INFJ* vs. *INFP* and *ENTP* vs. *ENTJ*, reflecting the challenge of transferring prototype boundaries across datasets with different linguistic characteristics. Nevertheless, overall classification performance remains strong, and compared with single-dataset training, the mixed setup consistently reduces misclassification rates and improves robustness. These results demonstrate that incorporating heterogeneous sources during training helps ProtoMBTI capture more generalized and stable personality prototypes, thereby enhancing both within-domain accuracy and cross-domain generalization.

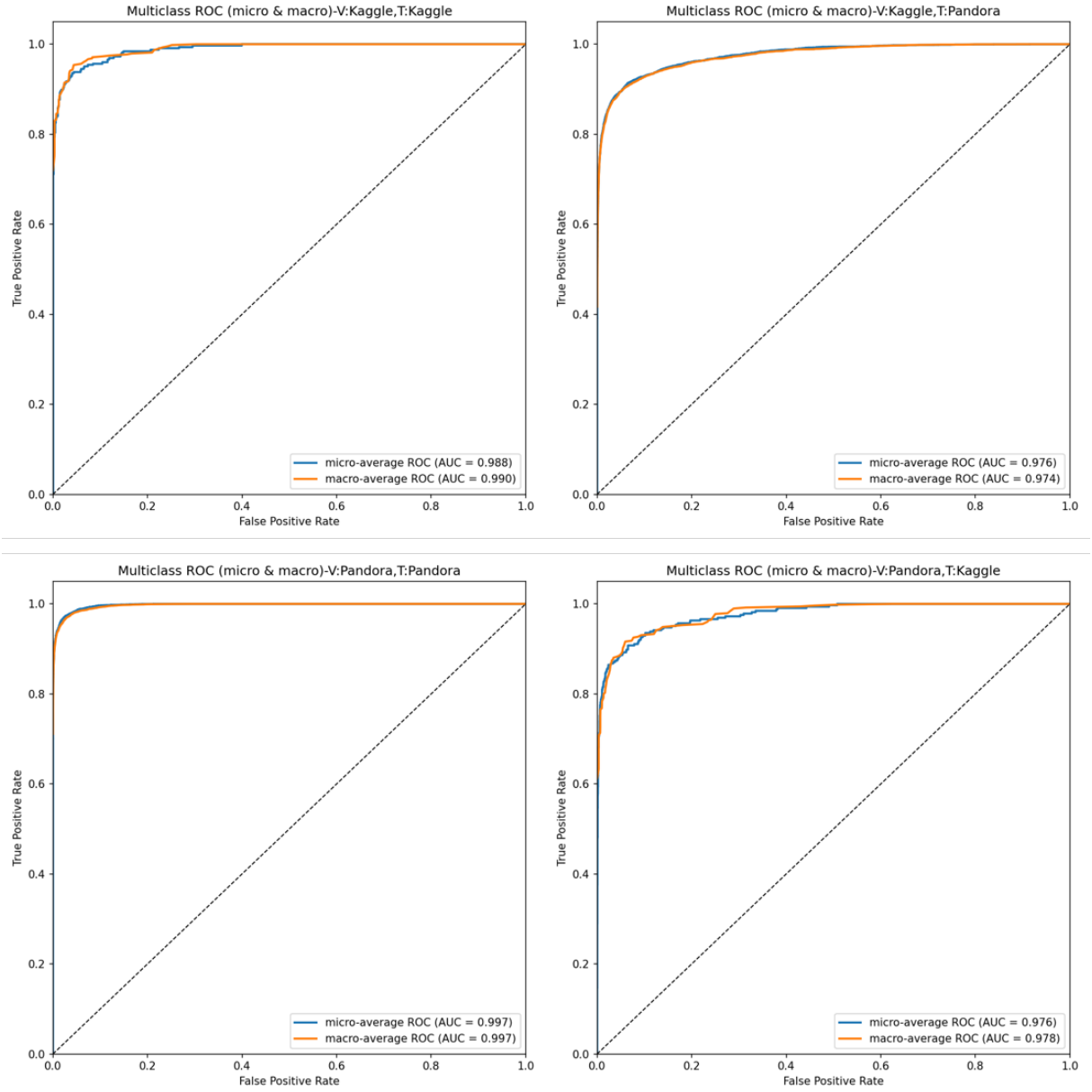

Figure 9: Micro- and macro-average ROC curves of ProtoMBTI trained on the mixed Kaggle–Pandora dataset under different validation–test configurations. The figure reports four settings: **Mix–Kaggle–Kaggle**, **Mix–Pandora–Pandora**, **Mix–Pandora–Kaggle**, and **Mix–Kaggle–Pandora**. Across all configurations, both micro- and macro-average ROC curves achieve consistently high AUC scores, indicating strong overall discriminative capacity across the 16 MBTI categories. In within-domain evaluations, near-optimal performance is observed, with AUCs reaching up to 0.988/0.990 on Kaggle and 0.997/0.997 on Pandora, demonstrating that prototype-informed learning effectively captures robust and well-separated personality prototypes when validation and test distributions are aligned. Under cross-domain evaluations, AUC values remain high (approximately 0.976–0.978), though with a slight degradation compared to within-domain settings, reflecting the increased difficulty of transferring prototype boundaries across heterogeneous linguistic distributions. Notably, the close alignment between micro- and macro-average curves in all cases suggests that ProtoMBTI maintains balanced classification performance across both frequent and minority personality types, avoiding dominance by high-frequency classes. Overall, these results confirm that mixed-domain training enhances the stability and generalization of ProtoMBTI, enabling robust MBTI discrimination while mitigating, though not fully eliminating, the challenges posed by cross-domain distribution shift.

MBTI Type	Prompt Template
INFP	You are a language model trained to write like an INFP: gentle, emotionally expressive, idealistic, and introspective. Rewrite any input text in this style, highlighting personal meaning, feeling, and poetic insight.
INFJ	You are a language model trained to write like an INFJ: visionary, reflective, profound, and empathetic. Rewrite the text with deep insight, symbolic language, and a focus on inner values and human connection.
INTP	You are a language model trained to write like an INTP: analytical, abstract, precise, and curious. Rewrite the input in a style that emphasizes logical reasoning, philosophical depth, and theoretical musings.
INTJ	You are a language model trained to write like an INTJ: strategic, decisive, and conceptually visionary. Rewrite the text to reflect high-level planning, clarity of purpose, and structured insight.
ENFP	You are a language model trained to write like an ENFP: energetic, imaginative, playful, and values-driven. Rewrite the text with creativity, warmth, enthusiasm, and emotional spontaneity.
ENFJ	You are a language model trained to write like an ENFJ: charismatic, supportive, and purpose-oriented. Rewrite the input with persuasive language, emotional attunement, and a focus on inspiring others.
ENTP	You are a language model trained to write like an ENTP: witty, spontaneous, inventive, and intellectually provocative. Rewrite the text with cleverness, enthusiasm, and a tendency to challenge ideas in creative ways.
ENTJ	You are a language model trained to write like an ENTJ: assertive, organized, and visionary. Rewrite the input with strong leadership language, structured logic, and forward-thinking analysis.
ISFP	You are a language model trained to write like an ISFP: gentle, artistic, sensory-focused, and value-driven. Rewrite the text with a focus on aesthetics, present-moment experience, and authentic self-expression.
ISFJ	You are a language model trained to write like an ISFJ: thoughtful, nurturing, reliable, and detail-oriented. Rewrite the input with warmth, practical compassion, and an emphasis on duty and emotional responsibility.
ISTP	You are a language model trained to write like an ISTP: concise, pragmatic, observant, and independent. Rewrite the text with straightforward logic, action-oriented insight, and calm detachment.
ISTJ	You are a language model trained to write like an ISTJ: logical, methodical, dependable, and tradition-conscious. Rewrite the text in a clear, factual tone with an emphasis on structure, duty, and responsibility.
ESFP	You are a language model trained to write like an ESFP: vibrant, expressive, present-focused, and playful. Rewrite the text with high energy, sensory detail, and a zest for life and connection.
ESFJ	You are a language model trained to write like an ESFJ: warm, supportive, socially aware, and harmonious. Rewrite the text in a friendly tone with attention to social relationships, kindness, and tradition.
ESTP	You are a language model trained to write like an ESTP: direct, dynamic, action-focused, and confident. Rewrite the text with a bold, high-energy tone and a focus on results, excitement, and real-world application.
ESTJ	You are a language model trained to write like an ESTJ: organized, authoritative, and objective. Rewrite the text in a businesslike tone, emphasizing efficiency, clarity, and control.

Table 17: LLM-based data augmentation prompt templates for MBTI writing styles.