Facet-Informed Prompting for LLM-Based Personality Assessment: Error-Guided Exemplar Selection and Hierarchical Prediction

Rasiq Hussain[1] **Juhi Chetan Shah**[1] **Joshua Oltmanns**[2] **Mehak Gupta**[1]
[1]Southern Methodist University [2]Washington University in St. Louis
{rasiqh, juhichetans, mehakg}@smu.edu j.oltmanns@wustl.edu

Abstract

Large language models (LLMs) are increasingly applied to automatic personality assessment, yet most prior work relies on coarse binary labels and direct domain-level predictions, limiting interpretability and ignoring the hierarchical facet structure of personality. In this study, we implement a structured prompting approach with three complementary objectives: direct domain-level prediction, fine-grained facet-level prediction, and domain-level prediction informed by facet outputs. All predictions use a five-level ordinal label scheme, capturing a continuum from very low to very high trait expression. Across all prompt types, we adopt an error-guided self-refinement procedure using in-context learning (ICL) to guide the model toward more accurate predictions. Zero-shot prompts assess baseline performance, while one-shot prompts incorporate a single demonstration example selected through the refinement procedure. Our framework evaluates both domain- and facet-level predictions, enabling examination of how prediction granularity and targeted exemplar selection influence LLM inference. By combining hierarchical domain-facet relationships with structured prompting and refinement, this work aims to provide a systematic approach for interpretable personality assessment from long-form life narratives, with broader implications for mental health research where fine-grained trait-level inference from speech can support clinical assessment.

1 Introduction

Personality is a central construct in psychology, shaping cognition, emotion, and behaviour. The Five-Factor Model (FFM) organizes personality into five broad domains—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (Costa Jr and McCrae, 1992)—each comprising six subordinate facets, operationalized in instruments such as the NEO-PI-R (Costa and McCrae, 2008). Facets capture unique trait variance, show only moderate intercorrelations, and exhibit heterogeneous behavioural and linguistic expressions (DeYoung et al., 2007; Soto and John, 2017), suggesting that predicting personality at the facet level before aggregating to domains may improve precision and interpretability.

Most computational approaches to personality assessment predict personality directly at the broad domain level using binary labels, classifying each trait simply as high or low (Mairesse et al., 2007; Majumder et al., 2017; Yang et al., 2023; Yeo et al., 2025). This simplified formulation overlooks two key aspects of the underlying psychometric model: the hierarchical facet structure within each domain and the continuous nature of trait expression. These limitations are particularly important in mental health research, where Big Five traits are linked to a wide range of psychological and behavioral outcomes (Ozer and Benet-Martinez, 2006; Kotov et al., 2010, 2017). As a result, fine-grained personality inference from naturalistic speech may support more scalable and interpretable assessment approaches for both clinical and research settings (Hussain et al., 2026; Le Glaz et al., 2021).

These gaps motivate three research questions guiding our investigation:

RQ1 Label resolution. How does increasing label resolution from binary to five-level ordinal classification affect LLM performance?

RQ2 Prediction granularity. Does a hierarchical approach, predicting at the facet level and aggregating to domain scores, yield more accurate domain-level assessments than direct domain-level prediction?

RQ3 Exemplar selection. Does semantically-guided retrieval of in-context exemplars targeting systematic errors improve performance relative to zero-shot and static one-shot prompting?

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 221–237

To address these questions, we developed a structured prompting framework consisting of three complementary prompt types: (i) direct domain-level prediction, (ii) facet-level prediction, and (iii) domain-level prediction conditioned on inferred facets. Each task uses a five-level ordinal scale (Very Low to Very High), requiring the model to capture subtle trait differences. Across all prompt types, we implement an error-guided self-refinement procedure via in-context learning (ICL), using task-relevant examples to improve predictions without updating model parameters (Dong et al., 2024; Liu et al., 2022; Madaan et al., 2023). Zero-shot prompts assess baseline performance, while one-shot prompts incorporate a single demonstration example selected through the refinement procedure. Facet-level predictions are used in prompt type III to inform domain-level scores, enabling comparisons of prediction granularity.

By combining fine-grained, five-level classification with structured facet-to-domain reasoning and targeted exemplar selection, this framework provides a psychometrically informed approach to evaluating and improving LLM-based personality assessment from long-form narratives.

2 Related Work

2.1 Large Language Models for Personality Assessment

The use of natural language processing for psychological and mental health assessment has gained increasing attention in recent years (Le Glaz et al., 2021; Sikström et al., 2025). Early work on personality prediction from text relied on handcrafted linguistic features and lexicons, demonstrating that stylistic and semantic cues can be predictive of personality traits (Mairesse et al., 2007; Argamon et al., 2007).

With the rise of neural models, representation learning approaches based on deep architectures further improved performance by capturing latent semantic patterns in text (Majumder et al., 2017). More recently, large language models (LLMs) have enabled a shift toward prompting-based approaches, where models infer personality traits directly from text without task-specific fine-tuning. Recent work has shown that LLMs can generate coherent personality assessments and capture personality-relevant signals in language, suggesting their potential for naturalistic personality inference (Peters et al., 2024; Rosenfelder et al., 2025; Zhu et al., 2025a; Hussain et al., 2026).

Despite these advances, most existing approaches focus on predicting personality at the broad domain level using coarse label spaces, often reducing trait prediction to binary or low-resolution classification (Mairesse et al., 2007; Majumder et al., 2017; Yang et al., 2023). This simplified formulation overlooks the continuous nature of trait expression. In contrast, our work adopts a five-level ordinal scheme to enable more precise and interpretable personality inference.

2.2 Domain-Guided Structured Prompting for Personality Inference

Recent work has explored incorporating psychological structure into prompt design to improve the reliability of LLM-based personality assessment. For example, Yang et al. (2023) proposes questionnaire-style prompting inspired by psychometric inventories, encouraging models to reason over structured trait definitions. PADO (Yeo et al., 2025) uses multiple LLM agents that analyze input text from distinct linguistic perspectives to infer relative levels of the Big Five (OCEAN) traits. Zhu et al. (2025b) propose that when you prompt an LLM to respond to individual items from a structured personality inventory (e.g., Big Five questionnaire items) rather than just asking for a direct overall trait prediction, the model produces more accurate personality inferences, even from short interview or text samples. These approaches show that grounding prompts in domain-specific frameworks can improve alignment with human judgments and provide more interpretable outputs.

In this work, we extend this direction by introducing a multi-stage prompting framework that first infers facet-level traits and then conditions domain-level predictions on these intermediate outputs.

2.3 Prompt Refinement for Personality Inference

Iterative self-refinement can improve performance by analyzing model outputs to identify and correct systematic errors (Madaan et al., 2023). Rather than relying on repeated calls to the model, refinement can also be implemented via in-context learning (ICL), which provides task-relevant demonstration examples directly in the prompt, allowing the model to adapt its predictions without updating parameters (Brown et al., 2020; Dong et al., 2024). ICL performance depends on the quality of demonstration examples, with dynamically selected se-

mantically relevant examples outperforming random or static choices (Liu et al., 2022; Rubin et al., 2022).

Building on these insights, we implement an error-guided prompt refinement strategy for personality inference. ICL demonstration examples are selected not only for semantic similarity but also to target systematic model errors, such as consistent mispredictions of specific traits or facets.

3 Data

3.1 SPAN: Life Narrative Interviews

This study uses Life Narrative Interview from St. Louis Personality and Aging Network (SPAN) study (Oltmanns et al., 2020). Participants narrate life experiences through family, work, and formative events in open-ended interviewer prompts. Personality was assessed using the Big Five framework, with each domain composed of six facets as defined in the NEO-PI-R (Costa and McCrae, 2008). Details about data processing are shared in Supplementary A.

Five-level label construction. Domain scores in the dataset range from 0 to 192. We standardize domain and facet scores in our dataset before assigning five-level labels. For each domain and facet, we compute a z-score $z = (x - \mu)/\sigma$ using the sample mean and standard deviation reported in Table 1. We then assign labels using z-score cutoffs motivated by standard psychometric practice for five-band trait interpretation: very low ($z < -1.5$), low ($-1.5 \leq z < -0.5$), average ($-0.5 \leq z < 0.5$), high ($0.5 \leq z < 1.5$), and very high ($z \geq 1.5$). These cutoffs divide the score distribution into five meaningful bands reflecting relative trait standing within the dataset, and are applied independently at both the domain and facet level.

4 Method

We design a structured prompting framework with three prompt types corresponding to different inference objectives: (1) direct domain-level prediction, (2) facet-level prediction, and (3) domain-level prediction derived from facet predictions.

We first describe the zero-shot setting to establish baseline model behavior under each prompt type. We then extend this setup to a one-shot setting, where a single demonstration example is incorporated into the prompt. Finally, we detail our refinement procedure for exemplar selection, which leverages error patterns observed in the zero-shot setting to guide the choice of informative in-context examples.

4.1 Prompt Types

4.1.1 Prompt Type I: Direct Domain-Level Prediction

The first prompt type performs direct prediction of the five personality domains (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) from the input text. A single prompt is used to jointly predict all five domains, with each domain assigned a label on a five-level ordinal scale (Very Low, Low, Average, High, Very High).

This prompt evaluates the model's ability to infer high-level personality traits directly from linguistic cues without explicit decomposition into facets. It serves as a baseline for comparison against more structured prediction strategies. Prompt I is implemented in both zero- and one-shot settings. Examples for each setting can be found in Supplementary B.

4.1.2 Prompt Type II: Facet-Level Prediction

The second prompt type performs fine-grained prediction at the facet level. For each personality domain, a dedicated prompt is used to predict its six underlying facets, resulting in six facet-level predictions per domain. Each facet is framed as a five-level ordinal classification task.

This decomposition allows the model to reason over more specific behavioral signals within the text. In addition, facet-level predictions provide a mechanism for analysing systematic errors and identifying which aspects of personality are most challenging for the model to infer. Prompt II is also implemented in both zero- and one-shot settings, and examples can be found in Supplementary C.

4.1.3 Prompt Type III: Domain-Level Prediction from Facets

The third prompt type predicts domain-level traits based on the facet-level outputs produced by Prompt Type II. This prompt takes as input the previously inferred facet-level labels and the inference (reasoning) for the corresponding prediction.

This formulation enables a structured reasoning step in which the model aggregates facet-level evidence to produce a final domain-level prediction. This approach allows the language model to learn implicit relationships between facets and domains

Table 1: Domain and facet-level dataset statistics showing mean, standard deviation, and transcript counts across five-level buckets.

Facet	Mean	Std	Very Low	Low	Average	High	Very High
Openness	113.1	18.7	94	401	621	347	113
Imagination	16.2	4.6	90	385	653	357	110
Artistic Interests	18.8	5.1	137	371	596	388	103
Emotionality	19.8	3.9	134	301	623	422	115
Adventurousness	16.0	3.9	129	448	459	432	127
Intellect	19.8	5.2	91	422	594	375	113
Liberalism	21.3	4.1	129	371	610	385	98
Conscientiousness	124.4	17.9	95	362	652	374	93
Self-Efficacy	23.4	3.3	115	297	807	268	108
Orderliness	17.7	4.1	122	295	651	424	103
Dutifulness	23.3	3.6	85	414	662	306	128
Achievement-Striving	19.0	4.1	107	318	722	374	75
Self-Discipline	21.0	4.5	147	253	696	418	81
Cautiousness	18.9	3.9	93	344	593	461	104
Extraversion	109.2	18.9	108	349	624	396	99
Friendliness	23.0	4.1	109	272	666	472	76
Gregariousness	16.7	5.0	134	368	617	399	77
Assertiveness	16.5	4.6	102	421	525	467	80
Activity Level	16.5	4.1	118	374	626	350	127
Excitement-Seeking	15.1	4.3	108	326	681	372	109
Cheerfulness	20.3	4.7	145	267	653	439	91
Agreeableness	130.9	16.1	99	347	672	335	123
Trust	21.9	4.2	118	286	557	563	71
Morality	22.6	4.1	121	332	620	414	108
Altruism	24.3	3.4	133	322	578	461	101
Cooperation	19.2	3.8	113	394	657	317	115
Modesty	20.0	4.1	99	282	793	334	87
Sympathy	21.8	3.4	93	439	570	375	117
Neuroticism	73.0	21.3	86	421	626	317	126
Anxiety	12.9	4.8	72	451	632	342	99
Anger	10.9	4.5	97	389	696	282	131
Depression	11.2	5.3	79	446	614	329	127
Self-Consciousness	13.2	4.2	61	539	568	295	132
Immoderation	15.1	4.2	79	363	712	324	117
Vulnerability	9.0	3.9	103	418	587	370	118

through contextual reasoning. Prompt III is identical for both zero-shot and one-shot settings because its primary role is to perform evidence aggregation from Prompt II, rather than generating predictions from raw transcripts itself. The example prompt is shared in the Supplementary D.

4.2 Zero-Shot Prompting

In the zero-shot setting, both Prompt I and Prompt II are executed without any labeled examples. This setup evaluates the model's inherent ability to perform personality inference across different levels of abstraction.

All zero-shot prompts follow a consistent structure using the standard *system* + *user* format. The system prompt defines the model's role and task, while the user prompt provides the input transcript. Each system prompt consists of the following components:(1) Expert Role Definition: The model is instructed to act as an expert personality psychologist with training in computational psycholinguistics, (2) Task Contextualization: For domain-level prompts (Prompt I), concise definitions of the five personality domains are provided. For facet-level prompts (Prompt II), definitions of the six facets within the target domain are included, (3) Analytic Directive: The model is instructed to analyze linguistic and behavioral cues in the text and assign ratings based on the provided definitions, (4) Reasoning Directive: The model is encouraged to justify each prediction using evidence from the text, promoting more grounded inferences, (5) Constrained Output Schema: Outputs are required in

a structured JSON format, including the predicted label, a brief rationale, and supporting evidence.

4.3 One-Shot Prompting with In-Context Refinement

To improve prediction accuracy, we extend both Prompt I and Prompt II with a one-shot configuration that incorporates a single demonstration example. This example is selected using the error-guided refinement procedure described in Section 4.4.

In the one-shot setting, the prompt retains the same structure as the zero-shot version but is augmented with two additional components: (1) Exemplar Demonstration: A labeled transcript is provided prior to the target input, including both the text and its corresponding domain or facet annotations, (2) Instructional Alignment: The model is explicitly instructed to use the demonstration as a reference for identifying relevant linguistic patterns and applying that knowledge when analyzing the target transcript.

We adopt a one-shot design to balance performance gains with prompt length constraints, particularly given the inclusion of long narrative transcripts and detailed facet definitions. Prior work has shown that excessive context length can reduce effective utilization of relevant information (Liu et al., 2024; Chang et al., 2024).

This unified prompting framework allows us to systematically compare: (i) direct vs. hierarchically structured prediction, (ii) domain vs. facet-level reasoning, and (iii) zero-shot vs. refinement-based ICL.

4.4 Error-Guided Refinement with In-Context Learning

We implement refinement through in-context learning (ICL). In our framework, the selection of demonstration examples is treated as a refinement decision. We first analyze where and how the model fails and then identify examples that target these failure modes. These examples are then incorporated into the prompt at inference time.

Error Analysis. The refinement loop begins with an automated error analysis on the training set. The model is first run in zero-shot mode. For Prompt I (domain-level prediction), misclassification is calculated at the domain level, while for Prompt II (facet-level prediction), it is calculated for each facet within a domain. Errors are decomposed into overestimation (predicting higher than ground truth) and underestimation (predicting lower than ground truth), as these require opposite corrective examples. A model that consistently overestimates a facet needs to be shown examples of low trait expression, while a model that underestimates needs examples of high trait expression. For Prompt I (domain-level prediction), we identify the two domains with the highest misclassification rates, and for Prompt II (facet-level prediction), we similarly identify the two facets with the highest misclassification rates per domain as primary targets for refinement. Selecting the top two targets balances focused error correction with generalizability. Improving these high-error domains or facets domains can also yield moderate gains in non-targeted domains or facets due to inter-domain and inter-facet correlations (Soto and John, 2017).

Example Pool and Directional Selection. To support the refinement process, we construct separate candidate example pools for Prompt I and Prompt II. For Prompt I, the pool contains training examples covering all label levels for each domain, while for Prompt II, the pool covers all label levels for each facet. Once the error analysis identifies the target domains or facets and their dominant error direction, we filter the respective pool accordingly. For overestimated targets, we retain only examples with ground-truth very-low or low labels; for underestimated targets, we retain only examples with ground-truth very-high or high labels. This directional filtering ensures that the selected demonstration is corrective for the specific failure mode identified.

Semantic Similarity Filtering. Directional filtering narrows the candidate set but does not guarantee that the retrieved demonstration is semantically aligned with the input transcript. To address this, we apply a second filtering step based on semantic similarity. All transcripts are embedded using OpenAI's `text-embedding-3-small` model, and cosine similarity is computed between the target transcript and candidate pool transcripts. The top-1 most similar transcript is selected for demonstration. Retrieval is implemented using FAISS for efficient nearest-neighbor search (Liu et al., 2022).

5 Experiment Design

All experiments use GPT-4o with temperature set to 0.0 for deterministic outputs. Each transcript is processed independently, and model outputs are

Table 2: Zero-shot Micro-F1 across three label resolutions - binary, three, and five-class resolution.

Domain	Binary	3-class	5-class
Openness	0.80	0.59	0.34
Conscientiousness	0.94	0.63	0.25
Extraversion	0.64	0.52	0.32
Agreeableness	0.87	0.70	0.28
Neuroticism	0.46	0.41	0.27

parsed to extract predicted labels.

The dataset is partitioned into disjoint training (50%) and test (50%) sets. The training set is used exclusively to construct the candidate pool for demonstration selection during in-context learning and for error analysis in the refinement procedure. The test set is used for final evaluation of each prompt performance, ensuring no data leakage.

We construct the example pool from the training partition of 788 transcripts. Specifically, all exemplar candidate transcripts are drawn exclusively from the training set, since no test transcript appears in the candidate pool at any point during the exemplar selection or error analysis. This ensures strict separation between evaluation and refinement.

To ensure sufficient coverage for the exemplar selection across all domains, facets, and label levels, we verify that every domain and facet–label-level combination is represented by at least 30 transcripts in the pool. Because every transcript carries labels for all 5 domains and 30 facets simultaneously, a single transcript contributes to multiple domains and facet-level.

We structure our evaluation around the three research questions from Section 1, examining the impact of label resolution, hierarchical facet-to-domain prediction, and error-guided in-context learning refinement. For each prompt type, we compare zero-shot and one-shot configurations, and analyze zero-shot errors to motivate the refinement process. We also conduct ablation studies to isolate the effects of error-guidance and contextual similarity in the refinement pipeline.

To complement Micro-F1 with a metric that reflects the ordinal structure of the label scheme, we additionally report ordinal tolerance accuracy. This metric measures the proportion of predictions that fall within one label of the true value, capturing near-miss performance that exact-match metrics do not reflect (Weber et al., 2025). For example, on a five-level scale, a prediction of High when the true label is Average represents a meaningfully smaller error than a prediction of Very Low, and ordinal tolerance accuracy captures this distinction in a way that Micro-F1 does not.

6 Results

6.1 Effect of Label Resolution

We first evaluate zero-shot performance across different label resolutions. Micro-F1 in Table 2 shows performance declines as the task shifts from binary to multi-level prediction. The largest relative drop between binary and 5-level is observed for Conscientiousness (73.4%), followed by Agreeableness (67.8%), Openness (57.5%), and Extraversion (50.0%). Neuroticism shows a comparatively smaller decrease (41.3%).

These results show that fine-grained personality prediction is substantially more challenging than coarse binary classification, motivating the need for structured prediction strategies at higher resolution.

6.2 Impact of Prediction Granularity

In Table 3, we examine whether modeling personality at the facet level improves domain-level prediction by comparing output of Prompt I and Prompt III.

Under zero-shot prompting, Prompt Type III improves over Prompt Type I across all domains, with gains for Openness (+8.8%), Conscientiousness (+28.0%), Extraversion (+18.8%), Agreeableness (+17.9%), and Neuroticism (+33.3%). This suggests that incorporating facet-level structure improves performance, though not uniformly.

Under one-shot prompting, the improvements are larger, with Prompt Type III outperforming Prompt Type I for Openness (+36.4%), Conscientiousness (+60%), Extraversion (+27.3%), Agreeableness (+33.3%), and Neuroticism (+34.5%), indicating that facet-level reasoning becomes more effective when combined with exemplar guidance.

6.3 Effect of Error-Guided Exemplar Selection

6.3.1 Zero-Shot Error Analysis

To guide exemplar selection for prompt refinement, we first examine systematic errors under zero-shot prompting (Table 4) to identify directional biases.

At the domain level, Openness, Conscientiousness, and Agreeableness are predominantly overestimated (76.3%, 90.7%, and 83.3% of errors, respectively), while Neuroticism is mainly underestimated (59.0%). Extraversion shows more balanced

Table 3: Domain-level Micro-F1 scores on test set for Prompt Type I (Direct Domain Prediction) and Prompt Type III (Domain from Facets). Bold indicates the best overall score per domain, while underlined values indicate the best score within Prompt Type I.

Domain	**Prompt Type I: Direct Domain**		**Prompt Type III: Domain from Facets**	
	Zero-Shot	One-Shot	Zero-Shot	One-Shot
Openness	0.34	0.33	0.37	**0.45**
Conscientiousness	0.25	0.25	0.32	**0.40**
Extraversion	0.32	0.33	0.38	**0.42**
Agreeableness	0.28	0.30	0.33	**0.40**
Neuroticism	0.27	0.29	0.36	**0.39**

errors (45.1% under vs. 54.9% over), indicating weaker directional bias.

Facet-level results closely mirror domain-level trends: facets under Openness, Conscientiousness, Extraversion, and Agreeableness are predominantly overestimated, while Neuroticism facets show strong underestimation. This alignment suggests that domain-level biases arise from systematic misclassification patterns originating at the facet level. We identify the top two contributing domains and facets per domain to implement error correction for Prompt I and Prompt II, respectively.

6.3.2 Performance Impact of Refinement

As observed in Tables 3 and 5, one-shot prompting generally improves performance over zero-shot at both domain and facet levels.

Domain-level improvements. At the domain level (Table 3), Prompt Type I shows only marginal gains from zero-shot to one-shot, within $\pm 7\%$ for most traits, indicating limited ability to leverage exemplars at a coarse domain level. In contrast, Prompt Type III, which aggregates facet-level information, achieves larger improvements: +21.6% for Openness, +25.0% for Conscientiousness, +10.5% for Extraversion, +21.2% for Agreeableness, and +8.3% for Neuroticism. These results highlight that facet-informed aggregation better translates exemplar guidance into domain-level gains.

Facet-level improvements. Targeting the top two most misclassified facets in each domain (Table 4) improves both the selected facets and non-targeted facets, demonstrating a clear "trickle-down" effect (Table 5).

In Openness, focusing on Emotionality (85.8% over) and Adventurousness (67.6% over) improves Micro-F1 by 20.4% and 19.8%, respectively, with additional gains of 4.7–6.7% for most other facets. For Conscientiousness, targeting Dutifulness (80.5% over) and Achievement-Striving (87.6% over) increases these facets by 29.9% and 31.4%, with smaller improvements of 5.1–16.2% in the remaining facets. In Extraversion, correcting Gregariousness (62.7% over) and Excitement-Seeking (63.6% over) yields approximately 5–14% gains across all facets. Agreeableness sees large improvements in Altruism (91.4% over, +30.7%) and Sympathy (91.0% over, +38.6%), with modest 7.7–13.3% gains for non-targeted facets. Finally, Neuroticism, targeting Anxiety (53.3% under) and Self-Consciousness (83% under), improves these facets by 13.8% and 24.8%, respectively, while other facets show smaller gains or slight decreases, reflecting weaker inter-facet correlations in this domain.

Translation to domain-level gains. These facet-level improvements directly contribute to the stronger domain-level performance of Prompt Type III, where zero- to one-shot gains range from 8–25%, compared with only 0–7% for Prompt Type I. This underscores how exemplar-guided, facet-informed prompting more effectively enhances overall predictions.

Exemplar guidance vs Random guidance. As an additional comparison, we evaluate a random one-shot baseline by randomly sampling one training example for each facet. The process was repeated 100 times and averaged for stability. The random baseline achieves an average facet-level Micro-F1 of 0.293, compared to 0.347 for the proposed error-guided one-shot strategy (Prompt Type II; Table 5). These results indicate that the observed improvements arise from targeted exemplar selection rather than merely including an arbitrary demonstration example.

6.4 Ordinal Tolerance Analysis

We additionally report ordinal tolerance accuracy for Prompt Type III domain-level predictions to complement Micro-F1 in Table 6. Tolerance accuracy evaluates model performance beyond exact-match accuracy, which measures the proportion

Table 4: Domain- and facet-level misclassification analysis on train set under zero-shot prompting. The table reports overall misclassification rate (%), along with the proportion of overestimation and underestimation errors expressed as percentages of total misclassifications (normalized such that Over + Under = 100%). The dominant error type per domain is shown in bold, and the top two most misclassified facets per domain are underlined.

Domain / Facet	Miss (%)	Over (%)	Under (%)
Openness	65.59	**76.30**	23.70
Imagination	61.58	40.56	**59.44**
Artistic Interests	60.51	**58.71**	41.29
Emotionality	71.97	**85.79**	14.21
Adventurousness	67.52	**67.64**	32.36
Intellect	61.25	**61.37**	38.63
Liberalism	60.36	**52.53**	47.47
Conscientiousness	74.67	**90.66**	9.34
Self-Efficacy	66.94	**82.15**	17.85
Orderliness	62.49	38.51	**61.49**
Dutifulness	71.72	**80.48**	19.52
Achievement-Striving	73.72	**87.61**	12.39
Self-Discipline	61.09	**63.30**	36.70
Cautiousness	63.64	42.87	**57.13**
Extraversion	67.33	**54.91**	45.09
Friendliness	57.87	**60.26**	39.74
Gregariousness	64.88	**62.65**	37.35
Assertiveness	60.84	**66.41**	33.59
Activity Level	62.41	**63.83**	36.17
Excitement-Seeking	66.47	**63.57**	36.43
Cheerfulness	60.26	**54.16**	45.84
Agreeableness	71.45	**83.30**	16.70
Trust	57.71	**66.15**	33.85
Morality	66.36	**69.94**	30.06
Altruism	77.16	**91.44**	8.56
Cooperation	64.17	**63.56**	36.44
Modesty	55.98	**60.39**	39.61
Sympathy	79.21	**91.02**	8.98
Neuroticism	72.92	41.03	**58.97**
Anxiety	64.09	46.67	**53.33**
Anger	62.50	31.44	**68.56**
Depression	63.16	**61.99**	38.01
Self-Consciousness	65.62	17.03	**82.97**
Immoderation	59.70	46.16	**53.84**
Vulnerability	63.52	45.41	**54.59**

Table 5: Facet-level Micro-F1 scores on test set for Prompt Type II: Facet-Level Prediction. Bold indicates the best score per facet.

Facet	Zero-Shot	One-Shot
Openness		
Imagination	0.387	**0.410**
Artistic Interests	0.401	**0.428**
Emotionality	0.275	**0.331**
Adventurousness	0.313	**0.375**
Intellect	**0.398**	0.391
Liberalism	0.402	**0.421**
Conscientiousness		
Self-Efficacy	0.315	**0.366**
Orderliness	0.385	**0.416**
Dutifulness	0.271	**0.352**
Achievement-Striving	0.245	**0.322**
Self-Discipline	0.373	**0.433**
Cautiousness	0.372	**0.391**
Extraversion		
Friendliness	0.409	**0.467**
Gregariousness	0.351	**0.401**
Assertiveness	0.383	**0.422**
Activity Level	0.376	**0.405**
Excitement-Seeking	0.334	**0.368**
Cheerfulness	0.409	**0.428**
Agreeableness		
Trust	0.428	**0.461**
Morality	**0.340**	0.308
Altruism	0.231	**0.302**
Cooperation	0.355	**0.396**
Modesty	0.421	**0.477**
Sympathy	0.207	**0.287**
Neuroticism		
Anxiety	0.355	**0.404**
Anger	**0.367**	0.270
Depression	**0.394**	0.391
Self-Consciousness	0.326	**0.407**
Immoderation	**0.417**	0.391
Vulnerability	0.360	**0.392**

of predictions falling within one label of the true value on the five-point scale. Unlike Micro-F1, this metric captures the clinical relevance of near-miss predictions, where a one-level discrepancy carries less consequence than a large ordinal error.

At the domain level, Prompt Type III achieves consistently high ordinal tolerance across all five domains under both zero-shot and one-shot prompting, with overall accuracies of 82.5% and 81.1%, respectively. This indicates that most domain-level predictions remain within one label of the ground truth regardless of prompting strategy.

The results also suggest a directional effect of error-guided exemplar selection. One-shot exemplars are intentionally selected to correct systematic overestimation or underestimation, which helps recover exact label matches and improves Micro-F1 across all domains. At the same time, this stronger corrective behavior can occasionally shift predictions farther from the true label for samples that were already within one adjacent category under zero-shot prompting, leading to a small reduction in ordinal tolerance. This trade-off is reflected in Table 3, where one-shot prompting improves exact classification performance while maintaining comparably high ordinal consistency in Table 6. Future work could explore softer ordinal-aware correction strategies that better balance exact-match accuracy and near-miss predictions.

6.5 Ablation Analysis

We performed an ablation study to isolate the contributions of error guidance and semantic similarity in exemplar selection for Prompt Type III (Ta-

Table 6: Ordinal tolerance accuracy (%) for Prompt Type III (Domain from Facets), reporting the proportion of domain-level predictions falling within one label of the true value under zero-shot and one-shot prompting.

Domain	Zero-Shot	One-Shot
Openness	86.4	84.7
Conscientiousness	81.6	77.9
Extraversion	83.5	81.1
Agreeableness	80.0	78.8
Neuroticism	80.8	83.1
Overall	82.5	81.1

Table 7: Ablation of exemplar selection. Domain-level Micro-F1 is shown for full one-shot (error + similarity), error-guided only, and similarity-guided only selection. Bold indicates the best score per domain.

Domain	Full One-Shot	Error Only	Similarity Only
Openness	**0.45**	0.40	0.42
Conscientiousness	**0.40**	0.37	0.38
Extraversion	**0.42**	0.37	0.39
Agreeableness	**0.40**	0.36	0.37
Neuroticism	**0.39**	0.34	0.34

ble 7). Using only error-guided exemplars, where examples are randomly selected from the subset of instances chosen based on error direction (i.e., selecting low or very low for overestimated facets and vice versa), reduces domain-level Micro-F1 by 7-12%, reflecting the inclusion of less relevant examples without error-guidance.

Selecting exemplars solely based on similarity, i.e., choosing the most semantically relevant examples without first filtering based on error direction, leads to smaller reductions per domain (5–12%) compared to error-guided exemplars.

Integrating both error guidance and similarity yields the best performance. This demonstrates that error guidance focuses the model on the most problematic facets, while similarity ensures the selected exemplars are contextually relevant. The synergy of these components explains why Prompt Type III outperforms other strategies.

7 Discussion

This study examined how prompting strategies, prediction granularity, and label resolution affect LLM-based personality assessment from life narratives. Our results demonstrate that incorporating hierarchical facet structure improves domain-level predictions compared with direct domain-level prompting. By modeling finer-grained facets, the model can capture subtle linguistic cues often missed at the coarse domain level. Higher label resolution (5-level) increases task difficulty, making facet-level modeling and exemplar guidance especially valuable.

Exemplar-based one-shot prompting is most effective at the facet-to-domain prompt, serving as targeted corrections for high-error facets. Improvements in these facets propagate to non-targeted facets within the same domain, mostly producing a positive "trickle-down" effect, though the magnitude varies depending on inter-facet correlations (DeYoung et al., 2007; Soto and John, 2017; Costa and McCrae, 2008). For example, domains such as Openness, Conscientiousness, and Agreeableness exhibit moderate correlations, resulting in mostly positive spillover effects, whereas Neuroticism facets are weakly correlated and show more heterogeneous improvements.

Structuring predictions hierarchically, first at the facet level and then aggregating to the domain level, allows exemplar-based corrections to translate into domain-level improvements. This approach provides practical guidance for applied LLM-based personality assessment, especially under multi-level label schemes.

The clinical relevance of this work extends beyond methodology. Big Five personality domains, particularly Neuroticism and Agreeableness, have established links to psychopathology, treatment adherence, and mental health outcomes(Kotov et al., 2010; Ozer and Benet-Martinez, 2006). Accurate fine-grained personality inference from life narratives could therefore support scalable mental health research by enabling automated trait assessment through text. While current performance levels position this framework as a research tool rather than a clinical instrument, the ordinal tolerance accuracy exceeding 82% at the domain level for Prompt Type III suggests meaningful signal capture. Validation against clinician-rated assessments represents a necessary next step toward clinical translation.

8 Conclusion

Incorporating facet-level prediction with error-guided exemplar refinement enhances LLM-based personality assessment, particularly at higher label resolutions. Aggregating facet predictions consistently improves domain-level performance and demonstrates the benefits of hierarchically structured prompting for more accurate and interpretable personality inference.

Ethics Statement

Personality assessment, particularly using automated tools, raises several ethical concerns. This study uses data from the SPAN study, which was approved by the Washington University in St. Louis Institutional Review Board. All transcripts are anonymized, personally identifiable information has been removed, and participants provided written informed consent as part of the original SPAN study protocol. We emphasize that LLM-based personality predictions should not be used as standalone clinical judgments. Instead, they are intended to provide supplementary insights, for example, in behavioral research or exploratory mental health studies. Misinterpretation or over-reliance on model outputs could lead to harmful consequences, particularly in sensitive contexts such as therapy, hiring, or legal settings. Researchers and practitioners should treat model outputs as probabilistic suggestions rather than definitive evaluations.

Limitations

While facet-based aggregation improves domain-level predictions, performance is still constrained by facets with abstract or sparsely expressed cues. Future work could explore decoupled facet-level prompting, where each facet is modeled independently before aggregation, potentially reducing error propagation and improving coverage of additional facets. Shorter-text sources such as social media posts may also provide complementary exemplars for more scalable refinement and out-of-distribution evaluation.

Another consideration is the label distribution in the dataset. Approximately 40% of transcripts receive an *average* label, while the *very low* and *very high* categories together account for only about 14% of observations, contributing to lower performance on extreme trait levels. Future work could improve robustness for these underrepresented categories through targeted sampling or cost-sensitive learning.

Finally, the ground-truth labels are derived from self-report NEO assessments rather than clinician-rated evaluations. Since self-perceived and externally observed traits may differ, further validation with clinician-rated personality assessments would strengthen the framework's clinical applicability. Nonetheless, the model demonstrates meaningful predictive performance despite this cross-modality setting, suggesting that naturalistic speech captures informative personality signals.

Acknowledgments

We thank the participants of the St. Louis Personality and Aging Network (SPAN) study for their contributions to this research. Computational resources were provided by the SMU O'Donnell Data Science and Research Computing Institute.

References

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Xiaoqian Liu, Tong Xiao, and Jingbo Zhu. 2024. Efficient prompting methods for large language models: A survey. *arXiv preprint arXiv:2404.01077*.

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

Paul T Costa Jr and Robert R McCrae. 1992. The five-factor model of personality and its relevance to personality disorders. *Journal of personality disorders*, 6(4):343–359.

Colin G DeYoung, Lena C Quilty, and Jordan B Peterson. 2007. Between facets and domains: 10 aspects of the big five. *Journal of personality and social psychology*, 93(5):880.

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

Rasiq Hussain, Zerui Ma, Ritik Khandelwal, Joshua Oltmanns, and Mehak Gupta. 2026. Language-based personality assessment from life narratives: a focus on model interpretability and efficiency. *Frontiers in Artificial Intelligence*, 9:1760246.

Roman Kotov, Wakiza Gamez, Frank Schmidt, and David Watson. 2010. Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. *Psychological bulletin*, 136(5):768.

Roman Kotov, Robert F Krueger, David Watson, Thomas M Achenbach, Robert R Althoff, R Michael Bagby, Timothy A Brown, William T Carpenter, Avshalom Caspi, Lee Anna Clark, and 1 others. 2017. The hierarchical taxonomy of psychopathology (hitop): A dimensional alternative to traditional nosologies. *Journal of abnormal psychology*, 126(4):454.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, and Christophe Lemey. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of medical Internet research*, 23(5):e15708.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, pages 100–114.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE intelligent systems*, 32(2):74–79.

Joshua R Oltmanns, Joshua J Jackson, and Thomas F Oltmanns. 2020. Personality change: Longitudinal self-other agreement and convergence with retrospective-reports. *Journal of Personality and Social Psychology*, 118(5):1065.

Daniel J Ozer and Veronica Benet-Martinez. 2006. Personality and the prediction of consequential outcomes. *Annu. Rev. Psychol.*, 57(1):401–421.

Heinrich Peters, Moran Cerf, and Sandra C Matz. 2024. Large language models can infer personality from free-form user interactions. *arXiv preprint arXiv:2405.13052*.

Ariel Rosenfelder, Maor Daniel Levitin, and Michael Gilead. 2025. Towards social superintelligence? ai infers diverse psychological traits from text without specific training, outperforming human judges. *Computers in Human Behavior: Artificial Humans*, page 100228.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 2655–2671.

Sverker Sikström, Ieva Valavičiūtė, and Petri Kajonius. 2025. Personality in just a few words: Assessment using natural language processing. *Personality and Individual Differences*, 238:113078.

Christopher J Soto and Oliver P John. 2017. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117.

Samantha Weber, Nicolas Deperrois, Robert Heun, Laura Frühschütz, Anna Monn, Stephanie Homan, Andrea Häfliger, Erich Seifritz, Tobias Kowatsch, MULTICAST consortium Jäger Lena 8 Schultebraucks Katharina 9 Gershov Sapir 9 Mocellin Jacopo 1 4, and 1 others. 2025. Using a fine-tuned large language model for symptom-based depression evaluation. *npj Digital Medicine*, 8(1):598.

Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. 2023. Psycot: Psychological questionnaire as powerful chain-of-thought for personality detection. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 3305–3320.

Haein Yeo, Taehyeong Noh, Seungwan Jin, and Kyungsik Han. 2025. Pado: Personality-induced multi-agents for detecting ocean in human-generated texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5719–5736.

Jianfeng Zhu, Ruoming Jin, and Karin G Coifman. 2025a. Can llms infer personality from real world conversations? *arXiv preprint arXiv:2507.14355*.

Jianfeng Zhu, Ruoming Jin, and Karin G Coifman. 2025b. Investigating large language models in inferring personality traits from user conversations. *arXiv preprint arXiv:2501.07532*.

A Data Preprocessing

Transcripts were anonymized, and initial acknowledgments were removed. Punctuation and casing were preserved to retain the natural structure of the narratives. Each transcript ranges from approximately 1,100 to 7,900 tokens, with an average

of 2,500 words. We removed 170 transcripts out of 1,408 total since they were either too short or too long, to avoid length-based bias in the experiments. We used the OpenAI text-embedding-3-small model with a context window of 8,191 tokens, ensuring that none of the transcripts were truncated.

B Prompt I Example

Figure 1 shows an example of a direct domain-level prompt in a zero-shot setting. Figure 2 shows an example of a direct domain-level prompt in a one-shot setting.

C Prompt II Example

Figure 3 shows an example of a facet-level prompt in a zero-shot setting. Figure 4 shows an example of a facet-level prompt in a one-shot setting.

D Prompt III Example

Figure 5 shows an example of Prompt III, which aggregates the evidence gathered by Prompt II to predict the domain-level label. This prompt is identical for both zero-shot and one-shot settings because its primary role is to perform evidence aggregation from Prompt II, rather than generating predictions from raw transcripts itself.

Prompt I: Zero-Shot Domain-Level Configuration

1. Expert Role Definition (System Prompt) You are an expert personality psychologist specializing in computational psycholinguistics. Your task is to evaluate linguistic evidence in text transcripts to infer personality domain levels for all five Big Five domains: *Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism*.

2. Task Contextualization (Trait Domain Definitions) Domain Definitions and Characteristics:

- **Openness:** Broad, complex, and imaginative mental life. Enjoys new experiences and abstract ideas.
- **Conscientiousness:** Organized, responsible, self-disciplined, and goal-oriented behavior.
- **Extraversion:** Energetic engagement with others, sociable, assertive, and positive affect.
- **Agreeableness:** Compassionate, cooperative, trusting, and considerate toward others.
- **Neuroticism:** Tendency toward emotional instability, anxiety, and negative affect.

3. Analysis and Reasoning Directive

1. Read the transcript carefully.
2. Identify linguistic, thematic, and behavioral cues relevant to each domain.
3. Assign a rating for all five domains.
4. Justify each rating using evidence from the transcript.

4. Rating Scale

- `very high`: strong evidence of high domain trait
- `high`: moderate evidence of high domain trait
- `average`: mixed evidence
- `low`: moderate evidence of low domain trait
- `very low`: strong evidence of low domain trait

5. Constrained Output Schema

```
{
 "Openness": {"rating": "...", "inference": "...", "quote": "..."},
 "Conscientiousness": {"rating": "...", "inference": "...", "quote": "..."},
 "Extraversion": {"rating": "...", "inference": "...", "quote": "..."},
 "Agreeableness": {"rating": "...", "inference": "...", "quote": "..."},
 "Neuroticism": {"rating": "...", "inference": "...", "quote": "..."}
}
```

6. Input Transcript (User Prompt)

```
[Insert transcript text here]
```

Figure 1: Zero-shot configuration for Prompt I, showing system role, domain definitions, analysis procedure, and output format.

Prompt I: One-Shot Domain-Level Configuration

1. Expert Role Definition (System Prompt) You are an expert personality psychologist specializing in computational psycholinguistics. Your task is to evaluate linguistic evidence in text transcripts to infer personality domain levels for all five Big Five domains: *Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism*.

2. Task Contextualization (Trait Domain Definitions) Domain Definitions and Characteristics:

- **Openness:** Broad, complex, and imaginative mental life. Enjoys new experiences and abstract ideas.
- **Conscientiousness:** Organized, responsible, self-disciplined, and goal-oriented behavior.
- **Extraversion:** Energetic engagement with others, sociable, assertive, and positive affect.
- **Agreeableness:** Compassionate, cooperative, trusting, and considerate toward others.
- **Neuroticism:** Tendency toward emotional instability, anxiety, and negative affect.

3. Analysis and Reasoning Directive

1. Read the transcript carefully.
2. Identify linguistic, thematic, and behavioral cues relevant to each domain.
3. Assign a rating for all five domains.
4. Justify each rating using evidence from the transcript.

4. Exemplar Demonstration

```
[Insert labeled transcript example with domain ratings here]
```

5. Instructional Alignment
Use the exemplar demonstration as a reference to guide your pattern recognition and ensure consistent reasoning when predicting domain-level ratings for the target transcript.

6. Rating Scale

- `very high`: strong evidence of high domain trait
- `high`: moderate evidence of high domain trait
- `average`: mixed evidence
- `low`: moderate evidence of low domain trait
- `very low`: strong evidence of low domain trait

7. Constrained Output Schema

```
{
 "Openness": {"rating": "...", "inference": "...", "quote": "..."},
 "Conscientiousness": {"rating": "...", "inference": "...", "quote": "..."},
 "Extraversion": {"rating": "...", "inference": "...", "quote": "..."},
 "Agreeableness": {"rating": "...", "inference": "...", "quote": "..."},
 "Neuroticism": {"rating": "...", "inference": "...", "quote": "..."}
}
```

8. Input Transcript (User Prompt)

```
[Insert target transcript text here]
```

Figure 2: One-shot configuration for Prompt I, showing system role, domain definitions, analysis procedure, exemplar, and output format.

Prompt II: Zero-Shot Configuration

1. Expert Role Definition (System Prompt) You are an expert personality psychologist specializing in computational psycholinguistics. Your task is to evaluate linguistic evidence in text transcripts to infer personality facet levels for *[Big Five domain, e.g., "Openness to Experience"]*.

2. Task Contextualization (Trait Facet Definitions) Trait Description: Openness reflects the breadth, depth, and complexity of an individual's mental and experiential life.

Facet Definitions and Characteristics:

- **Imagination (PNEOO1):**
  - **High:** Vivid imagination, enjoys fantasy and daydreaming.
  - **Low:** Concrete thinking, rarely engages in imaginative or abstract thought.
- **Artistic Interests (PNEOO2):**
  - **High:** Values and appreciates art, music, and aesthetic experiences.
  - **Low:** Avoids artistic activities, shows little interest in creative or cultural pursuits.
- **Emotionality (PNEOO3):**
  - **High:** Expresses emotions vividly and demonstrates empathy.
  - **Low:** Emotionally restrained, seldom reflects or communicates feelings.
- **Adventurousness (PNEOO4):**
  - **High:** Seeks novelty and variety, enjoys exploring new experiences.
  - **Low:** Prefers routine and familiar environments, avoids unfamiliar situations.
- **Intellect (PNEOO5):**
  - **High:** Enjoys abstract thinking, complex problem-solving, and analytical reasoning.
  - **Low:** Avoids complexity, prefers concrete or simple tasks.
- **Liberalism (PNEOO6):**
  - **High:** Endorses progressive values and openness to change.
  - **Low:** Holds traditional views, resists unconventional or progressive ideas.

3. Analysis and Reasoning Directive

1. Read the transcript carefully.
2. Identify linguistic and thematic cues.
3. Assign ratings for all six facets.
4. Justify each rating with inference and direct quote from the transcript.

4. Rating Scale

- `very high`: strong evidence of high trait
- `high`: moderate evidence of high trait
- `average`: mixed evidence
- `low`: moderate evidence of low trait
- `very low`: strong evidence of low trait

5. Constrained Output Schema

```
{
 "PNEOO1_scaled": {"rating": "...", "inference": "...", "quote": "..."},
 ...
 "PNEOO6_scaled": {...}
}
```

6. Input Transcript (User Prompt)

```
[Insert transcript text here]
```

Figure 3: Zero-shot configuration for Prompt II, showing system role, domain definitions, analysis procedure, and output format.

Prompt II: One-Shot Configuration

1. Expert Role Definition (System Prompt) You are an expert personality psychologist specializing in computational psycholinguistics. Your task is to evaluate linguistic evidence in text transcripts to infer personality facet levels for *[Big Five domain, e.g., "Openness to Experience"]*.

2. Task Contextualization (Trait Facet Definitions) Trait Description: Openness reflects the breadth, depth, and complexity of an individual's mental and experiential life.

Facet Definitions and Characteristics:

- **Imagination (PNEOO1):**
  - **High:** Vivid imagination, enjoys fantasy and daydreaming.
  - **Low:** Concrete thinking, rarely engages in imaginative or abstract thought.
- **Artistic Interests (PNEOO2):**
  - **High:** Values and appreciates art, music, and aesthetic experiences.
  - **Low:** Avoids artistic activities, shows little interest in creative or cultural pursuits.
- **Emotionality (PNEOO3):**
  - **High:** Expresses emotions vividly and demonstrates empathy.
  - **Low:** Emotionally restrained, seldom reflects or communicates feelings.
- **Adventurousness (PNEOO4):**
  - **High:** Seeks novelty and variety, enjoys exploring new experiences.
  - **Low:** Prefers routine and familiar environments, avoids unfamiliar situations.
- **Intellect (PNEOO5):**
  - **High:** Enjoys abstract thinking, complex problem-solving, and analytical reasoning.
  - **Low:** Avoids complexity, prefers concrete or simple tasks.
- **Liberalism (PNEOO6):**
  - **High:** Endorses progressive values and openness to change.
  - **Low:** Holds traditional views, resists unconventional or progressive ideas.

3. Analysis and Reasoning Directive

1. Read the transcript carefully.
2. Identify linguistic and thematic cues.
3. Assign ratings for all six facets.
4. Justify each rating with inference and direct quote from the transcript.

4. Exemplar Demonstration

`[Insert labeled transcript example with domain ratings here]`

5. Instructional Alignment

Use the exemplar demonstration as a reference to guide your pattern recognition and ensure consistent reasoning when predicting domain-level ratings for the target transcript.

6. Rating Scale

- `very high`: strong evidence of high trait
- `high`: moderate evidence of high trait
- `average`: mixed evidence
- `low`: moderate evidence of low trait
- `very low`: strong evidence of low trait

7. Constrained Output Schema

```
{
 "PNEOO1_scaled": {"rating": "...", "inference": "...", "quote": "..."},
 ...
 "PNEOO6_scaled": {...}
}
```

8. Input Transcript (User Prompt)

`[Insert target transcript text here]`

Figure 4: One-shot configuration for Prompt II, showing system role, domain definitions, analysis procedure, exemplar, and output format.

Prompt Type III: Facet-to-Domain Aggregation Example

INPUT: Facet-Level Ratings with Evidence
Each facet is assigned a rating along with an evidence-based inference extracted from the transcript.

Imagination	*Rating: High* *Inference: The speaker frequently engages in abstract thinking and discusses hypothetical scenarios, indicating strong imaginative capacity.*
Artistic Interests	*Rating: Average* *Inference: Limited references to artistic activities; occasional appreciation but not a dominant theme.*
Emotionality	*Rating: High* *Inference: The speaker expresses emotions vividly and reflects on internal experiences in detail.*
Adventurousness	*Rating: Low* *Inference: Preference for routine and avoidance of unfamiliar situations are repeatedly mentioned.*
Intellect	*Rating: High* *Inference: Strong engagement with complex ideas and analytical reasoning is evident throughout the narrative.*
Liberalism	*Rating: Average* *Inference: Some openness to new perspectives is present, but views remain moderately conventional.*

TASK: Domain-Level Aggregation
Using the facet ratings and their inferences above, determine a single domain-level rating for **Openness to Experience**.

GUIDELINES:

1. Consider the overall pattern across all six facets.
2. Assess whether the evidence consistently supports the high or low pole, or is mixed.
3. Resolve conflicts by prioritizing the inference text over the rating labels.

OUTPUT:
Openness to Experience: High
Justification: The majority of facets (Imagination, Emotionality, and Intellect) show strong high-pole evidence, outweighing the lower Adventurousness score and mixed signals in Artistic Interests and Liberalism. Overall, the narrative reflects a consistent pattern of cognitive exploration and emotional depth, supporting a high domain-level rating.

Figure 5: Prompt III, facet inputs and instructions to aggregate evidence and labels to give domain-level prediction, and output format.