Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 173-183 recibido 05-04-2025 revisado 06-05-2025 aceptado 18-05-2025

Personality Assessment on Spanish and Basque Texts using In-Context Learning Techniques

Evaluación de la Personalidad a partir de Textos en Español y Euskera mediante Técnicas de Aprendizaje en Contexto

Aitzol Saizar, Maddalen Lopez de Lacalle, Xabier Saralegi

Orai NLP Technologies
{m.lopezdelacalle,x.saralegi}@orai.eus

Abstract: This study assesses the performance of Llama3 generative large language models (8B and 70B) in predicting Big Five personality traits from Spanish and Basque texts. Various in-context learning approaches, including zero-shot, few-shot, and Chain-of-Thought (CoT) prompting, as well as instruction fine-tuning, were evaluated on two datasets built on texts from different sources, Essays and PAN-15 (with a Basque subset translated for this work). Results show that Llama3 performs poorly in Basque, with in-context learning strategies failing to exceed the random baseline, except for a slight improvement with CoT on the 70B model. Fine-tuning the 8B model provides only marginal gains. Performance in Spanish is better but remains modest, with one-shot prompting and fine-tuning offering slight improvements in the case of the smaller model. Finally, in the case of Spanish, all in-context learning techniques surpass zero-shot when using the 70B model.
Keywords: Personality recognition, Large Language Models, In-Context Learning, Low-resource Languages.

Resumen: Este trabajo evalúa el rendimiento de los modelos de lenguaje generativos Llama3 (8B y 70B) en la predicción de los rasgos de personalidad del modelo Big Five a partir de textos en español y euskera. Se analizaron diversas estrategias de aprendizaje en contexto, incluyendo zero-shot, few-shot y Chain-of-Thought (CoT) prompting, así como el ajuste fino, utilizando dos conjuntos de datos construidos a partir de diferentes fuentes: Essays y PAN-15 (con un subconjunto en euskera traducido específicamente para este trabajo). Los resultados muestran que Llama3 tiene un desempeño deficiente en euskera; las estrategias de aprendizaje en contexto no logran superar la línea de base aleatoria, salvo una leve mejora con CoT en el modelo de 70B. El ajuste fino del modelo de 8B solo proporciona mejoras marginales. En español, el rendimiento es superior pero sigue siendo modesto, el one-shot prompting y el ajuste fino ofrecen ligeras mejoras en el caso del modelo más pequeño. Por último, en el caso del español, todas las técnicas de aprendizaje en contexto superan el enfoque zero-shot cuando se utiliza el modelo de 70B.
Palabras clave: Evaluación de la personalidad, Modelos de Lenguaje de Gran Escala, Aprendizaje en Contexto, Lenguas con pocos recursos.

1 Introduction

Automatic text personality assessment has gained significant interest in recent years due to its potential applications in diverse fields such as human resources, mental health, education, and marketing (Mehta et al., 2020; Stachl et al., 2020). In an era where digital communication dominates, vast amounts of text data are generated daily, offering valuable insight into individual traits. Using natural language processing (NLP) for personality assessment can enhance user experiences in various domains, such as personalized marketing, recruitment, and mental health diagnostics. For instance, companies can tailor content to align with user personalities, improving engagement, while in HR, it facilitates candidate screening by assessing com-

ISSN 1135-5948 DOI 10.26342/2025-75-13

patibility with team dynamics. Additionally, in therapy or coaching, personality profiling can support mental health professionals in offering more targeted and effective interventions (Le Glaz et al., 2021; Calvo et al., 2017).

Text-Based Automatic Personality Recognition identifies an individual's personality traits through their written language by leveraging NLP and machine learning techniques. These methods extract meaningful linguistic features, such as word choice, syntax, sentiment, and language patterns, from text (Guo, 2022; Kazameini et al., 2020; Ren et al., 2021) and map them to established psychological models such as the Big Five traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (McCrae and John, 1992). While considerable progress has been made in this field, research has primarily focused on English, leaving a gap in studies exploring personality recognition in other languages. This work addresses that gap by focusing specifically on Spanish and Basque texts, extending the scope of multilingual personality assessment.

Large Language Models (LLMs) have already transformed many fields, leading to an increasing amount of research exploring their utility in automatic personality assessment (Yang et al., 2023). In this study, we evaluated the effectiveness of open-source Llama3 LLM in its 8B and 70B variants to assess Big Five personality traits from text. We compare various approaches, including zero-shot, few-shot and CoT in-context learning techniques, and fine-tuning the models on domain-specific datasets. The experiments focus on the two official languages of the Basque Country and Navarre—Spanish and Basque—using two distinct datasets with texts from diverse sources, including social networks and personal essays, for evaluation.

The research questions addressed in this article are as follows:

- RQ1: Is the zero-shot prompting strategy feasible for predicting personality traits for Spanish and Basque?
- RQ2: Do few-shot and chain-of-thought (CoT) prompting strategies improve the ability of the LLM to assess personality traits compared to a zero-shot approach?
- RQ3: Does fine-tuning offer significant performance improvements over in-context learning for personality recognition, considering the higher data and computational costs associated with fine-tuning?
- RQ4: How does the performance of Llama's smaller model (8B) compare to that of the larger model (70B) in personality trait prediction?

From here on, the paper is structured as follows. Section 2 reviews related works. Section 3 describes the dataset used in this work. Next, in Section 4 we detail the experimental setup. The results obtained are described in Sections 5. Finally, Section 6 draws conclusions on the experiments carried out.

2 Related Work

Early approaches to automatic personality recognition from text primarily relied on traditional machine learning techniques that utilized handcrafted linguistic features, such as lexical, syntactic, and semantic cues. Mairesse et al. (2007) demonstrated that specific linguistic features, including word choice and syntactic patterns, were correlated with the Big Five personality traits in both spoken and written language, laying the groundwork for text-based personality assessment.

With the rise of deep learning, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been effectively employed to automatically learn and capture relevant patterns from large text corpora. Majumder et al. (2017) introduced a CNN-based architecture to identify high-level personality-related features, outperforming traditional models by capturing more complex linguistic nuances. Similarly, recurrent models such as LSTMs and GRUs have been utilized to capture temporal dependencies in text, further enhancing personality trait prediction (Gjurković and Šnajder, 2018). Lynn, Balasubramanian, and Schwartz (2020) extended this line of research by proposing a hierarchical BiLSTM model with message-level attention to better capture personality cues from sequences of social media posts.

Subsequently, Transformer-based models, particularly BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), further advanced personality recognition by allowing fine-tuning with minimal labeled data. Mehta et al. (2020) showed that BERT, when

fine-tuned on personality datasets, surpassed previous models in both accuracy and consistency.

The advent of LLMs has further advanced the field, especially when combined with in-context learning techniques (Ke et al., 2024). Recent studies, such as PsyCoT (Yang et al., 2023) and DesPrompt (Wen et al., 2023), leverage LLMs in multi-turn conversational formats or fill-in-the-blank tasks to capture intricate linguistic nuances and draw more precise personality inferences. For example, PsyCoT facilitates gradual personality profiling through iterative questioning, while DesPrompt generates predictions based on descriptive adjective-driven prompts, proving particularly effective in low-data scenarios. However, most existing research remains focused on English texts, highlighting the need for further studies on multilingual personality recognition.

3 Datasets

The datasets used in this study contain text samples annotated with trait values, which we binarize for classification. Both datasets contain self-reported Big Five personality trait labels. Participants completed a self-report Big Five personality inventory, rating themselves on a series of statements. These responses were then scored to generate the personality trait labels. To ensure diversity in textual sources, we selected two datasets: one consisting of social media posts and the other comprising personal essays. Specifically, the following widely used datasets for text-based personality assessment were chosen for our experiments:

1. **Essays** (Pennebaker and King, 1999): This dataset consists of personal essays written by individuals, reflecting on their daily lives and personal thoughts, who have also completed the Big Five personality assessment questionnaire. Each text is labeled with scores and binary labels for the Big Five personality traits, allowing researchers to analyze and predict personality based on natural, self-expressive writing. Comprises 2468 students English essays labeled with the writers' Big-Five personality traits. For our experiments, we randomly selected a subset of 100 essays, along with their corresponding labels, to construct the test set (Essays_en), striving to maintain a balanced label distribution (see Table 2). This sample size was chosen to ensure a manageable evaluation process while still providing sufficient diversity in writing styles and personality traits. Additionally, after the subset was machine translated into Spanish, it was manually reviewed to create the Spanish version of the test set (Essays_es). This manual review process, essential for ensuring accuracy and quality, would have been impractical to scale across the entire dataset due to limitations in time and resources.

2. **PAN-AP-2015** (Rangel Pardo et al., 2015): The PAN-AP-2015 corpus data is sourced from Twitter and includes collections of tweets from individuals in four languages: English, Spanish, Italian, and German. In this study, we focus on the English texts (294 individuals in total: 152 in the training set and 142 in the test set) and Spanish texts (188 individuals: 88 in the training set and 100 in the test set). Each individual has approximately 100 tweets, with Big Five personality traits scored on a scale from -0.5 to +0.5. For binary classification, we categorized scores above 0.1 as "high", while all other scores were labeled as "low". This threshold was selected following a grid search strategy inspired by (Wen et al., 2023), in which multiple cutoff points within the normalized range [-0.5, +0.5] were evaluated. The value of 0.1 was chosen as it provided the most balanced distribution of positive (high) and negative (low) samples across traits, helping to mitigate class imbalance in the binary classification setup. In our experiments, we used the complete original test sets in English and Spanish (PAN_en and PAN_es, hereinafter). For the Basque experiments, we randomly selected 26 samples from the Spanish test set and manually translated them into Basque, creating the PAN_eu dataset[1]. This dataset holds particular significance as it represents the first personality recognition dataset available for the Basque language.

[1] huggingface.co/datasets/orai-nlp/PAN15-eu

Table 1 presents the datasets used in our experiments, along with their respective sizes for the test split. Each number indicates the number of individuals in the test set (i.e., documents written by distinct participants).

Dataset	Test
Essays_en	100
Essays_es	100
PAN_en	142
PAN_es	88
PAN_eu	26

Table 1: Test sets used for the experiments and the number of examples in each.

Both datasets are annotated using the Big-Five model, commonly referred to as the OCEAN model (McCrae and John, 1992). According to this model, personality is described along five dimensions:

- **Openness:** Imaginative, curious, and open to new experiences.
- **Conscientiousness:** Organized, reliable, and responsible.
- **Extraversion:** Sociable, talkative, and energetic.
- **Agreeableness:** Compassionate, cooperative, and kind.
- **Neuroticism:** Prone to emotional instability, anxiety, and moodiness.

It is important to note that the PAN-AP-2015 dataset labels Neuroticism inversely, as "emotional stability". We adapted the annotations by assigning a Boolean value of 0 for Neuroticism to individuals with high emotional stability, and vice versa.

Table 2 provides a summary of the distribution between the two categories for each trait across all datasets.

4 Experimental Setup

4.1 Models

For our experiments, we employed Meta's open-weight Llama3 multilingual LLMs (Dubey et al., 2024) in two configurations: 8 billion (8B) and 70 billion (70B) parameters. These models were chosen for their accessibility, allowing for extensive experimentation without the financial constraints of proprietary models, and for their strong multilingual capabilities, despite being primarily trained on English and other major languages. By leveraging both the smaller (8B) and larger (70B) models, we sought to explore how model size influences the accuracy of Big Five personality trait inference using zero-shot, few-shot, and CoT prompting strategies.

4.2 Prompting Strategies for personality assessment

4.2.1 Zero-Shot Prompting

For the zero-shot prompting approach for assessing personality traits, we designed a prompt that includes the instructions to guide the instructed LLM towards the specific task, along with the text from which the value of a specific personality trait should be inferred. The instruction is written in English, while the content text is in the source language –Spanish or Basque– depending on the language in which the traits are intended to be inferred. The specific contents of the prompt are shown in Table 3.

The placeholder [text] is replaced with the context content used to determine the value (high or low) of the specified trait, indicated by the placeholder [trait]. Furthermore, the model is requested to produce the output in a predefined format to facilitate automatic evaluation.

4.2.2 Few-Shot Prompting

In the few-shot prompting experiments, we designed a prompt that includes labeled examples of personality assessment task. To enhance the model's understanding of the task, we include a set of trait assessment examples, adjusting the number of examples based on the experimental setup (1-shot, 2-shot, or 4-shot), where x-shot refers to x labeled examples per trait. All examples were randomly sampled from the training set. Accordingly, the [number] placeholder is replaced with one, two, or four to align with the respective configuration. Following this structure, in the example provided in Table 6 (Appendix A), the placeholder [Text_1] is substituted with the first example's content, and [Text_2] with the second example's content. Similarly, the [text] placeholder is replaced with the contextual input used to evaluate the trait (specified by [trait]) as either high or low.

To study the impact of varying the number and diversity of the examples included in the prompt, we tested the following setups:

Dataset (%)	Opn	Con	Ext	Agr	Neu
PAN_en	67 / 33	51 / 49	51 / 49	51 / 49	49 / 51
PAN_es	52 / 48	46 / 54	53 / 47	66 / 34	52 / 48
PAN_eu	42 / 58	42 / 58	54 / 46	69 / 31	50 / 50
Essays_en	53 / 47	46 / 54	49 / 51	54 / 46	44 / 56
Essays_es	53 / 47	46 / 54	49 / 51	54 / 46	44 / 56

Table 2: Percentage distribution of positive (left) and negative (right) labels per Big Five trait across datasets.

```
You are an AI assistant who
specializes in text analysis.
You will complete a text analysis
task. The task is as follows:
according to a text written by an
author, predicting whether
the author is A:"High [trait]"
or B:"Low [trait]".

AUTHORS TEXT:
[text]

Write a choice in the format:
"CHOICE: "
and do not give the explanation.
```

Table 3: Zero-shot prompt used to infer the personality trait value.

- **1-shot:** We examine the effect of a single polarized example, where all instances are labeled with the same psychological trait, either high or low.
- **2-shot:** We assess the model's performance using two examples, with one example labeled high and the other low, ensuring a balanced representation of both classes.
- **4-shot:** We expand the prompt to include four examples, with two examples for each class, representing both high and low labels equally.

4.2.3 Chain-of-Thought (CoT) Prompting

For the Chain-of-Thought (CoT) prompting strategy, we used the prompt from (Yang et al., 2023), detailed in Appendix B. This strategy guides the model to evaluate personality-related statements step-by-step, scoring them on a scale from 1 to 5 based on the provided text (see Table 7). Each psychological trait has a specific set of statements, specified by the placeholder [range_statements], which the model evaluates sequentially. The scoring scale is as follows:

- 1: Strongly Disagree
- 2: Disagree a Little
- 3: Neutral
- 4: Agree a Little
- 5: Strongly Agree

After scoring all statements, the model is informed of which ones positively or negatively influence the trait, specified by the placeholder [statements_ids] and is then prompted to determine whether the trait is high or low based on the cumulative scores.

The general prompt used in the CoT strategy is shown in Table 7 in Appendix B. As noted, the placeholder [range_statement] is replaced with the range of statement IDs to be evaluated (e.g., S0-S9). Similarly, the [trait] and [text] placeholders are substituted with the specific trait being evaluated and the context used to determine the score for the statements, respectively. Finally, the [statements_ids] placeholder is replaced with the set of statement IDs that have a positive or negative association with the trait.

Table 8 in Appendix B illustrates the loop used to score all trait-specific statements. Here, the [statement_id] placeholder is replaced with the identifier of the statement (e.g., S0), while the [statement] placeholder holds the sentence of the assertion to be scored on a scale of 1 to 5 (e.g, The author always likes to collaborate with others). The LLM generates the response in the [score] placeholder.

Lastly, Table 9 in Appendix B provides the prompt used to elicit the LLM to assess the trait as high or low, considering the scores

of the statements, which may have either a positive or negative influence on the overall assessment.

4.3 Instruction Fine-Tuning

To adapt the LLaMA3-8B model to the task of personality assessment, we performed instruction fine-tuning using two complementary datasets: the Essays dataset (Pennebaker and King, 1999), which contains reflective, formal English texts, and the PAN-AP-2015 dataset (Rangel Pardo et al., 2015), which includes informal social media posts in both English and Spanish. This bilingual and stylistically diverse corpus aims to improve the model's ability to generalize across a broad spectrum of linguistic contexts and communication styles.

Instruction fine-tuning was conducted using Low-Rank Adaptation (LoRA) (Hu et al., 2022), a parameter-efficient fine-tuning method that enables scalable adaptation of large language models without modifying all model weights. The fine-tuning process employed a batch size of 64, 4 epoch and a cosine learning rate scheduler with a peak learning rate of 2e-5. The LoRA-specific hyperparameters were set to a rank of 64, an alpha value of 16, and a dropout probability of 0.1.

Training was carried out using NVIDIA RTX A5000 GPUs, with memory-efficient techniques such as 8-bit optimizers and gradient check-pointing enabled to accommodate the model within available hardware. For the training framework, we relied on the Hugging Face Transformers library (Wolf et al., 2020), integrating it with DeepSpeed (Rajbhandari et al., 2020) to handle memory partitioning and scale across multiple GPUs when necessary.

This stage of fine-tuning aimed to steer the base LLaMA3 model toward a more instruction-following behavior, aligning its outputs with the requirements of the personality assessment task. By exposing the model to high-level instructions and diverse input types during training, we expected to improve its robustness and reasoning in downstream evaluation, particularly in zero-shot scenarios.

5 Results

This section presents the experimental results of the instructed LLM Llama3 (8B and 70B) on the personality assessment task using zero-shot, few-shot, and CoT prompting approaches, as well as fine-tuning. The evaluation is conducted on the Essays and PAN-AP-2015 datasets across English, Spanish, and Basque.

We report results in terms of accuracy, measured as the proportion of correctly classified instances for each of the five Big Five traits. The overall accuracy is calculated as the average accuracy across all traits.

5.1 In-Context Learning Evaluation

As shown in the table 4, the zero-shot approach shows a poor performance in general, making it an unfeasible strategy for personality assessment. Its performance is particularly bad in the case of Basque, where it does not exceed the random baseline. The results for English and Spanish remain comparable, with accuracy scores of 0.56 and 0.54 in the Essays dataset and 0.55 and 0.52 in the PAN_15 dataset, respectively. Furthermore, no significant performance differences were observed between the two model sizes.

According to the results obtained for the few-shot prompting systems, the one-shot approach outperforms the zero-shot approach for Spanish, particularly when using the larger Llama model (70B) in both datasets. This improvement is most pronounced in the Essays dataset, where one-shot prompting with Llama3-70B yields the highest performance for Spanish. This is not the case for Basque, as the one-shot strategy does not outperform the zero-shot in most configurations. Adding more examples to the prompt did not improve the results, as neither the two-shot nor four-shot approaches outperformed the one-shot strategy for Spanish and English. For Basque, the four-shot approach only surpassed the one-shot strategy when using the larger model. Finally, the CoT strategy outperforms all other prompt-based approaches on the PAN-15 dataset when using the largest model for both Spanish and Basque. However, this is not observed in the Essays dataset for Spanish, where the best results are achieved with one-shot prompting.

Among all prompt-based approaches, the one-shot strategy achieves the best results on the Essays dataset, regardless of model size. In the PAN-15 dataset, one-shot also performs best with the smaller model, but the larger model enables CoT to surpass it.

Method	Essays_en		Essays_es		PAN_en		PAN_es		PAN_eu	
	8B	70B	8B	70B	8B	70B	8B	70B	8B	70B
Random	0.50		0.50		0.50		0.50		0.50	
Zero-shot	0.55	0.56	0.55	0.54	0.53	0.55	0.51	0.52	0.47	0.48
One-shot_pos	0.56	0.57	0.52	0.57	**0.54**	**0.57**	0.53	0.54	0.46	0.47
One-shot_neg	**0.58**	**0.59**	**0.57**	**0.59**	0.50	0.53	0.48	0.53	0.49	0.48
Two-shot	0.54	0.58	0.56	0.56	0.48	0.50	0.51	0.55	0.45	0.41
Four Shot	0.54	0.55	0.55	0.56	0.51	0.55	0.52	0.54	0.46	0.50
CoT	**0.58**	0.58	0.55	0.58	0.48	0.55	0.47	**0.57**	0.48	**0.52**
IFT	**0.58**	-	**0.57**	-	**0.54**	-	**0.54**	-	**0.52**	-

Table 4: Results for all prompt-based systems and the fine-tuning approach evaluated on the Essays and PAN-15 datasets.

However, it is important to note that CoT prompting is significantly more computationally expensive than one-shot.

5.2 Fine-Tuning Evaluation

Instruction fine-tuning of the Llama3-8B model leads to further performance improvements on the PAN-AP-2015 dataset, especially in Spanish and Basque. This confirms the effectiveness of supervised learning in capturing trait-specific patterns, especially in low-resource languages like Basque, where fine-tuning substantially improves performance compared to zero-shot and few-shot approaches (e.g., 0.52 vs. 0.47 average accuracy on PAN-15). For Basque fine-tuning is the only strategy that enables Llama3-8B to outperform the random baseline, highlighting the challenges of few-shot and zero-shot in-context learning strategies in low-resource settings.

However, the results differ notably for the Essays dataset. In English, fine-tuning achieves accuracy equivalent to the best-performing prompt-based methods (0.58). For Spanish the fine-tuned model scores 0.57, slightly below the top prompt-based result (0.59 by One-shot_neg with Llama3-70B). This suggests that while fine-tuning provides consistent improvements across languages, its relative advantage over prompt-based methods diminishes in high-resource scenarios where large language models can effectively leverage in-context learning.

5.3 Trait-wise Evaluation

To gain deeper insight into the model's behavior, we conducted a trait-wise performance analysis. Table 5 present the accuracy obtained per trait for Llama's 8B model across multiple prompting strategies and both datasets.

The effectiveness of trait classification methods varies by language. Fine-tuning generally provides the best performance, especially for Openness, Extraversion, and Neuroticism, though improvements are often modest. Prompting strategies show inconsistent benefits—CoT notably enhances Agreeableness in Basque (0.69) but underperforms for Extraversion in English (0.39). One-shot negative prompting is effective for Conscientiousness and Extraversion in Basque, while one-shot positive prompting improves Conscientiousness and Neuroticism in Spanish. Zero-shot surprisingly performs best for Openness in English (0.67). These results highlight the language-dependent effects of prompting and the limited yet stable advantage of fine-tuning.

The breakdown of the results for each trait in the Essays dataset (see Table 5) show that performance improves consistently with more advanced strategies for Openness, where fine-tuning achieves the highest accuracy on both English (0.61) and Spanish (0.63). Conscientiousness remains relatively stable across strategies, with fine-tuning performing best in English (0.57), while Spanish results fluctuate without notable improvements over zero-shot. Extraversion proves the most challenging trait to classify, with the weakest overall performance, especially in Spanish, where CoT drops to 0.45. Agreeableness maintains relatively high accuracy across all strategies, with CoT performing best in both languages, though improvements over zero-shot are modest. Neuroticism remains stable, with zero-shot matching the highest accuracy across strategies in both

PAN-15 Dataset							
Language	**Strategy**	**OPN**	**CON**	**EXT**	**AGR**	**NEU**	**AVG**
EU	Zero-Shot	0.38	0.54	0.58	0.42	0.42	0.47
	One-Shot_pos	0.42	0.38	0.54	0.46	0.50	0.46
	One-Shot_neg	0.46	0.62	0.62	0.31	0.46	0.49
	CoT	0.38	0.38	0.46	0.69	0.46	0.48
	Finetuning	0.46	0.50	0.50	0.58	0.54	0.52
EN	Zero-Shot	0.67	0.51	0.52	0.47	0.48	0.53
	One-Shot_pos	0.65	0.59	0.51	0.48	0.44	0.53
	One-Shot_neg	0.55	0.51	0.50	0.49	0.41	0.49
	CoT	0.60	0.51	0.39	0.47	0.45	0.48
	Finetuning	0.56	0.59	0.53	0.48	0.53	0.54
ES	Zero-Shot	0.46	0.52	0.54	0.56	0.49	0.51
	One-Shot_pos	0.46	0.68	0.49	0.45	0.59	0.53
	One-Shot_neg	0.48	0.31	0.58	0.45	0.58	0.48
	CoT	0.33	0.68	0.51	0.44	0.41	0.47
	Finetuning	0.57	0.57	0.55	0.53	0.50	0.54
Essays Dataset							
Language	**Strategy**	**OPN**	**CON**	**EXT**	**AGR**	**NEU**	**AVG**
EN	Zero-Shot	0.53	0.51	0.51	0.58	0.63	0.55
	One-Shot_pos	0.54	0.55	0.53	0.58	0.63	0.57
	One-Shot_neg	0.60	0.55	0.55	0.55	0.61	0.57
	CoT	0.56	0.55	0.58	0.59	0.61	0.58
	Finetuning	0.61	0.57	0.55	0.57	0.58	0.58
ES	Zero-Shot	0.53	0.55	0.48	0.60	0.58	0.55
	One-Shot_pos	0.52	0.51	0.50	0.59	0.50	0.52
	One-Shot_neg	0.61	0.54	0.51	0.62	0.56	0.57
	CoT	0.62	0.51	0.45	0.63	0.56	0.55
	Finetuning	0.63	0.55	0.56	0.56	0.57	0.57

Table 5: Accuracy by personality trait for each strategy on PAN-15 and Essays datasets. Traits: Openness (OPN), Conscientiousness (CON), Extraversion (EXT), Agreeableness (AGR), Neuroticism (NEU).

languages.

Fine-tuning consistently achieves the best or near-best results, particularly for the most challenging traits—those where zero-shot performance is weakest—demonstrating its effectiveness. One-shot negative prompting tends to outperform its positive counterpart, emphasizing the influence of example framing on model performance. CoT yields mixed results, excelling in certain traits (e.g., Agreeableness in Spanish) while underperforming in others (e.g., Extraversion in Spanish). Overall, Extraversion proves to be the most difficult trait to classify, whereas Agreeableness and Neuroticism exhibit relatively stable performance across strategies.

6 Conclusions

The primary objective of this study was to evaluate the performance of the Llama3-instructed LLMs (8B and 70B) in assessing the Big Five personality traits in Basque and Spanish texts. To achieve this, we explored several in-context learning approaches—zero-shot, few-shot, and CoT prompting—and fine-tuned the models using domain-specific instructions. Two widely used datasets, Essays and PAN-15, grounded in the Big Five psychological model, served as the basis for our experiments.

Our findings indicate that both Llama3 models (8B and 70B) struggle to perform the personality assessment task for Basque and Spanish in a zero-shot prompting setup. Per-

formance in Basque, in particular, was especially poor, failing to exceed the random baseline (0.47 for Llama3-8B and 0.48 for Llama3-70B). In truth, only the fine-tuning of the small model and the most advanced prompting strategy, CoT, have managed to modestly surpass the baseline performance.

While few-shot and CoT prompting approaches yielded modest improvements over the zero-shot baseline for Spanish, the improvements were relatively limited. Among all prompt-based strategies, the one-shot approach achieved the best results on the Essays dataset, for both model sizes. On the PAN-15 dataset, the one-shot strategy also outperformed others for the smaller model, but the larger model allowed CoT prompting to achieve slightly better results.

Fine-tuning the LLMs with domain-specific instructions resulted in only marginal improvements on the PAN-15 dataset for Spanish and Basque, compared to the prompting-based strategies.

Our experiments reveal that assessing personality traits from text is a challenging task for LLMs. Notably, model size did not lead to significant differences in performance, as both the 8B and 70B models produced similarly low results. This highlights the need for further research into leveraging LLMs effectively for psychological assessment tasks.

Acknowledgments

This work has been partially funded by the Basque Government (ICL4LANG project, grant no. KK-2023/00094). Model training and development were conducted using the Hyperion system at the Donostia International Physics Center (DIPC).

References

Calvo, R. A., D. N. Milne, M. S. Hussain, and H. Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dubey, A., A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Rozière, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. M. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Gjurković, M. and J. Šnajder. 2018. Reddit: A gold mine for personality prediction. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 87–97.

Guo, J. 2022. Deep learning approach to text analysis for human emotion detection from big data. *Journal of Intelligent Systems*, 31(1):113–126.

Hu, E. J., S. Yelong, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kazameini, A., S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria. 2020. Personality trait detection using bagged

svm over bert word embedding ensembles. *arXiv preprint arXiv:2010.01309*.

Ke, L., S. Tong, P. Cheng, and K. Peng. 2024. Exploring the frontiers of llms in psychological applications: A comprehensive review. *arXiv preprint arXiv:2401.01519*.

Le Glaz, A., Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. Devylder, M. Walter, S. Berrouiguet, et al. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of medical Internet research*, 23(5):e15708.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach, Sep.

Lynn, V., N. Balasubramanian, and H. A. Schwartz. 2020. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5306–5316.

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

Majumder, N., S. Poria, A. Gelbukh, and E. Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE intelligent systems*, 32(2):74–79.

McCrae, R. R. and O. P. John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Mehta, Y., N. Majumder, A. Gelbukh, and E. Cambria. 2020. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, 53(4):2313–2339.

Pennebaker, J. W. and L. A. King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Rajbhandari, S., J. Rasley, O. Ruwase, and Y. He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Rangel Pardo, F. M., F. Celli, P. Rosso, M. Potthast, B. Stein, and W. Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. In *CLEF 2015 evaluation labs and workshop working notes papers*, pages 1–8.

Ren, Z., Q. Shen, X. Diao, and H. Xu. 2021. A sentiment-aware deep learning approach for personality detection from text. *Information Processing & Management*, 58(3):102532.

Stachl, C., F. Pargent, S. Hilbert, G. M. Harari, R. Schoedel, S. Vaid, S. D. Gosling, and M. Bühner. 2020. Personality research and assessment in the era of machine learning. *European Journal of Personality*, 34(5):613–631.

Wen, Z., J. Cao, Y. Yang, H. Wang, R. Yang, and S. Liu. 2023. Desprompt: Personality-descriptive prompt tuning for few-shot personality recognition. *Information Processing & Management*, 60(5):103422.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Yang, T., T. Shi, F. Wan, X. Quan, Q. Wang, B. Wu, and J. Wu. 2023. PsyCoT: Psychological questionnaire as powerful chain-of-thought for personality detection. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3305–3320, Singapore, December. Association for Computational Linguistics.

A Appendix 1: Few-shot prompt

Table 6 details the few-shot prompt for personality trait classification, including task in-

structions, two labeled examples, and the target text.

```
You are an AI assistant who
specializes in text analysis.
You will complete a text analysis
task. The task is as follows:
according to a text written by an
author, predicting whether
the author is A:"High [trait]"
or B:"Low [trait]".
[number] examples:
[text_1]
CHOICE: A
[text_2]
CHOICE: B
AUTHORS TEXT:
[text]
Write a choice in the format:
"CHOICE: "
and do not give the explanation.
```

Table 6: Few-shot prompt for predicting a personality trait from the author's text.

B Appendix 2: Chain-of-Though prompt

This appendix presents the Chain-of-Thought (CoT) prompting strategy used to guide the model through a step-by-step trait assessment process. The approach involves a multi-turn dialogue where the model rates a series of trait-related statements based on an author's text before making a final high/low classification.

Tables 7–9 show the full structure of the CoT prompt.

```
According to the above scores
and the text, the author is more
likely to be: A: "High [trait]" or
B: "Low [trait]". Provide a choice
in the format: "CHOICE: <A/B>" and
do not give the explanation.

Assistant: CHOICE: [choice]
```

Table 9: Final prompt used in the CoT strategy to evaluate the personality trait value (high or low).

```
You are an AI assistant who
specializes in text analysis and
I am Human. We will complete a
text analysis task together through
a multi-turn dialogue.
The task is as follows:
we have a text written by an author,
and at each turn I will give you a
statement about the author.
According to the author's text,
you need to rate the statement with
a score from 1-5, where:
1 = Disagree strongly,
2 = Disagree a little,
3 = Neutral,
4 = Agree a little,
5 = Agree strongly.
After rating all the statements
([range_statements]), I will ask
you if the author is more likely
to be A: "High [trait]" or
B: "Low [trait]",
and then you need to give your
choice. Note that [statements_ids]
are positive statements,
with higher scores indicate
higher [trait], while
[statements_ids] are reverse-scored
statements, with higher scores
indicate lower [trait].

AUTHOR'S TEXT: [text]
```

Table 7: CoT prompt used to guide the LLM through the trait assessment process.

```
[statement_id]: "[statement]".
Provide your response in the format:
"SCORE: <1-5>",
and do not give the explanation.

Assistant: [score]
```

Table 8: Prompt containing the loop used in the CoT strategy, to evaluate personality-related statements for specific traits.