IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/45AI/2025/4538925/
Vol. 2025 (20), Article ID 4538925

How Well Do the Most Popular Large Language Models (LLMs) Recognize Myers-Briggs Type Indicator (MBTI) Personality Traits? *

Joanna WOZNIAK

Military University of Technology, Warsaw, Poland

Correspondence should be addressed to: Joanna WOZNIAK, joanna.wozniak@wat.edu.pl

* Presented at the 45th IBIMA International Conference, 25-26 June 2025, Cordoba, Spain

Abstract

This study evaluates the effectiveness of three popular large language models (LLMs)—GPT-3.5-Turbo, GPT-4o-Mini, and Gemini-2.0-Flash—in identifying Myers-Briggs Type Indicator (MBTI) personality traits from textual data. Using a dataset of over 8,000 personality-labeled posts, each model's capability to accurately classify individual MBTI letters (Introversion-Extraversion, Sensing-Intuition, Thinking-Feeling, Judging-Perceiving) and the full four-letter personality types was assessed. The Gemini-2.0-Flash model consistently demonstrated superior performance across all metrics, achieving the highest exact-match accuracy (72%) and balanced F1-scores. In contrast, GPT-3.5-Turbo showed strong biases toward majority classes, especially in distinguishing between intuitive and sensing traits, while GPT-4o-Mini presented marked improvements but remained less effective overall compared to Gemini-2.0-Flash. The results underscore the importance of class-sensitive metrics such as recall and F1-score, highlighting that high accuracy alone can mask significant model biases, especially within imbalanced datasets. Future research should focus on class balancing, fine-tuning for fairness, and extending the analysis to alternative personality frameworks and diverse text sources.

Keywords: LLM, MBTI, GPT, Gemini

‘Introduction

In recent years, large language models (LLMs) have gained widespread recognition, even among the general public, largely due to the success of OpenAI’s ChatGPT. This class of models includes systems such as ChatGPT, Gemini, DeepSeek, and LLaMA. LLMs can make mistakes and aren’t made for mathematical problems but are great tool for searching information, programming, assisting in decision making, pattern identification, analyzing data and many more[1].

Personality recognition is valuable in numerous real-world contexts. Its applications include adaptive e-learning systems that tailor content based on individual learning styles[2], as well as personality-based candidate evaluation in recruitment processes, where data such as resumes or social media activity is analyzed to match individuals to suitable job positions [3], [4]. Another emerging research direction in the field of personality modeling involves embedding consistent personality traits into AI systems, aiming to facilitate coherent and natural human–AI interaction dynamics [5].

Cite this Article as: Joanna WOZNIAK, Vol. 2025 (20) “How Well Do the Most Popular Large Language Models (LLMs) Recognize Myers-Briggs Type Indicator (MBTI) Personality Traits? " Communications of International Proceedings, Vol. 2025 (20), Article ID 4538925, https://doi.org/10.5171/2025.4538925

Myers-Briggs Personality Type (MBTI)

According to MBTI, personality classification is based on interaction between four basic preferences which are dichotomies: Extraversion or Introversion (E-I), Sensing or Intuition (S-N), Thinking or Feeling (T-F), Judging or Perceiving (J-P). [6] Those dichotomies are summarized in Table 1.

Table 1: The Four Dichotomies of the MBTI[6]

Extraversion-Introversion Dichotomy (attitudes or orientations of energy)	
Extraversion (E)	Introversion (I)
Directing energy mainly toward the inner world of experiences	Directing energy mainly toward the outer world of people and objects and ideas
Sensing-Intuition Dichotomy (functions or processes of perception)	
Sensing (S)	Intuition (N)
Focusing mainly on perceiving patterns and interrelationships	Focusing mainly on what can be perceived by the five senses
Thinking-Feeling Dichotomy (functions or processes of judging)	
Thinking (T)	Feeling (F)
Basing conclusions on logical analysis with a focus on objectivity and detachment	Basing conclusions on personal or social values with a focus on understanding and harmony
Judging-Perceiving Dichotomy (attitudes or orientations toward dealing with the outside world)	
Judging (J)	Perceiving (P)
Preferring the decisiveness and closure that result from dealing with the outer world using one of the Judging processes (Thinking or Feeling)	Preferring the flexibility and spontaneity that results from dealing with the outer world using one of the Perceiving processes (Sensing or Intuition)

Those interactions result in 16 personalities types represented by four letters determining the preferred poles in example: ESTJ would be personality characterized by extraversion, sensing, thinking and judging. All personalities are shown in Table 2.

Table 2. Contributions Made by Each Preference to Each Type[6]

		Sensing Types		**Intuitive Types**	
		With Thinking	**With Feeling**	**With Feeling**	**With Thinking**
Introverts	**Judging Types**	**ISTJ** **I** Depth of concentration **S** Reliance on facts **T** Logic and analysis **J** Organization	**ISFJ** **I** Depth of concentration **S** Reliance on facts **F** Warmth and sympathy **J** Organization	**INFJ** **I** Depth of concentration **N** Grasp of possibilities **F** Warmth and sympathy **J** Organization	**INTJ** **I** Depth of concentration **N** Grasp of possibilities **T** Logic and analysis **J** Organization
	Perceiving Types	**ISTP** **I** Depth of concentration **S** Reliance on facts **T** Logic and analysis **P** Adaptability	**ISFP** **I** Depth of concentration **S** Reliance on facts **F** Warmth and sympathy **P** Adaptability	**INFP** **I** Depth of concentration **N** Grasp of possibilities **F** Warmth and sympathy **P** Adaptability	**INTP** **I** Depth of concentration **N** Grasp of possibilities **T** Logic and analysis **P** Adaptability

Extraverts	**Perceiving Types**	**ESTP** **E** Breadth of interests **S** Reliance on facts **T** Logic and analysis **P** Adaptability	**ESFP** **E** Breadth of interests **S** Reliance on facts **F** Warmth and sympathy **P** Adaptability	**ENFP** **E** Breadth of interests **N** Grasp of possibilities **F** Warmth and sympathy **P** Adaptability	**ENTP** **E** Breadth of interests **N** Grasp of possibilities **T** Logic and analysis **P** Adaptability
	Judging Types	**ESTJ** **E** Breadth of interests **S** Reliance on facts **T** Logic and analysis **J** Organization	**ESFJ** **E** Breadth of interests **S** Reliance on facts **F** Warmth and sympathy **J** Organization	**ENFJ** **E** Breadth of interests **N** Grasp of possibilities **F** Warmth and sympathy **J** Organization	**ENTJ** **E** Breadth of interests **N** Grasp of possibilities **T** Logic and analysis **J** Organization

Related Work

The goal of this study is to evaluate and compare the performance of three prominent LLMs — GPT-3.5-Turbo, Gemini 2.0 Flash, and GPT-4o-Mini — in the task of MBTI personality classification. Specifically, we assess each model's ability to correctly classify individual letters of the personality type and analyze their overall performance in predicting the full four-character sequence.

Several studies have explored more traditional approaches to the Myers-Briggs Type Indicator (MBTI) personality recognition using Natural Language Processing (NLP) including data-centric approach [7] and models based on Bidirectional Encoder Representations from Transformers (BERT)[8]. Recent studies have explored how ChatGPT can recognize MBTI personality traits without explicit training [9], and how newer models such as Qwen2-7B and Llama3-8B perform in direct comparison [10].

Method

To compare effectiveness of ChatGPT and Gemini in assigning Myers-Briggs Personality Types to user-generated posts, we used a dataset containing posts and MBTI labels of authors [11]. We compared three models: gpt-3.5-turbo, gemini-2.0-flash and gpt-4o-mini. Those models were prompted with the query:

Analyze the following text and determine the author's MBTI personality type (e.g., INTP, ESFJ). Respond ONLY with the four-letter MBTI type and nothing else. Text: {POSTS}

The models responded with MBTI personality (for example INTP) for 8294 records in dataset. To evaluate model performance, we calculated standard classification metrics which are: accuracy, precision, recall, F1-score. Additionally, we generated confusion matrices for each letter in the MBTI personality type and computed exact-match accuracy over the full four-letter sequence.

Results

Model performance metrics resulting from the evaluation are summarized in Table 3. As shown in Figure 1, the Google's gemini-2.0-flash model achieved the highest F1-score values. This means the model has a good balance between precision and recall and has high utility value.

Table 3. Summary metrics

	Model	Letter	Accuracy	Precision	Recall	F1

1	gpt-3.5-turbo	I/E	0.889197	0.777591	0.726702	0.751286
2	gpt-3.5-turbo	N/S	0.914878	0.884752	0.437719	0.585681
3	gpt-3.5-turbo	F/T	0.867615	0.877899	0.826190	0.851260
4	gpt-3.5-turbo	J/P	0.806004	0.795280	0.912172	0.849724
5	gpt-4o-mini	I/E	0.907885	0.810738	0.782723	0.796484
6	gpt-4o-mini	N/S	0.928864	0.875683	0.562281	0.684829
7	gpt-4o-mini	F/T	0.885339	0.890257	0.855377	0.872469
8	gpt-4o-mini	J/P	0.839161	0.803658	0.969320	0.878749
9	gemini-2.0-flash	I/E	0.913190	0.814815	0.806283	0.810526
10	gemini-2.0-flash	N/S	0.940439	0.844350	0.694737	0.762271
11	gemini-2.0-flash	F/T	0.889800	0.846154	0.928478	0.885406
12	gemini-2.0-flash	J/P	0.845310	0.806724	0.976740	0.883628

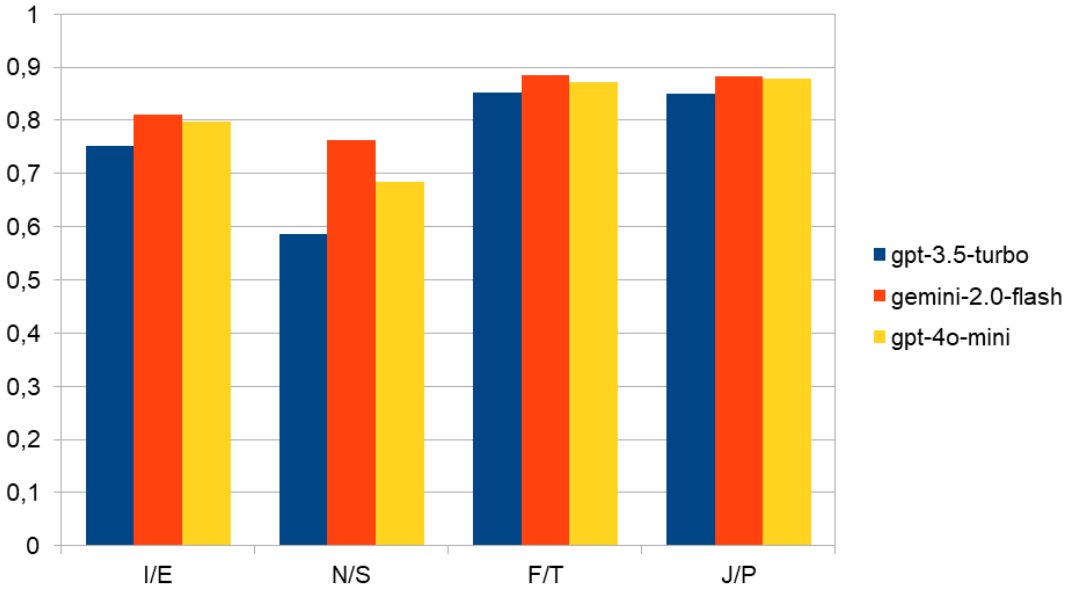

Figure 1: F1-score comparison for each model and letter

All models were evaluated in prediction of each set of letters of personality type using standard classification metrics. We present subset of results and confusion metrics that demonstrate characteristic patterns or significant differences between the models, such as imbalances between classes or variations across models.

Table 4 shows that gemini-2.0-flash achieves high precision and recall for both classes (for introversion F1=0.94 and for extraversion F1 = 0.81). This indicates good ability to generalize and effective differentiation between introverted and extroverted traits. Gemini-2.0-flash also maintains a good balance between classes, which is reflected in the high weighted average (F1 = 0.91).

Table 4: Classification report – gemini-2.0-flash – letter I/E

	precision	**recall**	**f1-score**	**support**
Introversion (I)	0.94	0.95	0.94	6384
Extraversion (E)	0.81	0.81	0.81	1910
macro avg	0.88	0.88	0.88	8294
weighted avg	0.91	0.91	0.91	8294

According to Table 5, despite the high precision for both classes (0.92 and 0.88), gpt-3.5-turbo shows very low sensitivity (recall = 0.44) for the sensing class, which results in a low F1-score (F1 = 0.59). This indicates a

potential bias toward the majority class (intuition), exemplifying how high overall accuracy (91%) can obscure poor performance in the minority class.

Table 5: Classification report – gpt-3.5-turbo – letter N/S

	precision	recall	f1-score	support
Intuition (N)	0.92	0.99	0.95	7154
Sensing (S)	0.88	0.44	0.59	1140
macro avg	0.9	0.71	0.77	8294
weighted avg	0.91	0.91	0.9	8294

As shown in Table 6, gemini-2.0-flash demonstrates strong and balanced performance on both classes, achieving balanced precision and recall. The F1-score is 0.89 for both the feeling and thinking classes, making this dichotomy one of the most reliably modeled in the entire experiment. This result suggests that emotional-logical differences are easier to capture by the model in the linguistic context.

Table 6: Classification report – gemini-2.0-flash – letter F/T

	precision	recall	f1-score	support
Feeling (F)	0.93	0.86	0.89	4491
Thinking (T)	0.85	0.93	0.89	3803
macro avg	0.89	0.89	0.89	8294
weighted avg	0.89	0.89	0.89	8294

The results in Table 7 indicate that gpt-4o-mini achieves high sensitivity for the perceiving class (recall = 0.97), which translates into strong recognition of the perceptual (P) trait. At the same time, a notably lower recall for the judging class (recall = 0.64) results in asymmetric prediction. Despite this imbalance, the overall F1-score (0.88 for the perceiving class, 0.76 for the judging class) indicates model's overall effectiveness in distinguishing between the two traits.

Table 7: Classification report – gpt-4o-mini – letter J/P

	precision	recall	f1-score	support
Judging (J)	0.93	0.64	0.76	3307
Perceiving (P)	0.8	0.97	0.88	4987
macro avg	0.87	0.81	0.82	8294
weighted avg	00.86	0.84	0.83	8294

We have chosen to show confusion matrices that are the most standing out per every letter in MBTI personality type. For letter I/E we have selected confusion matrix for gemini-2.0-flash model which is shown in Table 8 which exhibited the best F1-score for this letter (F1 = 0.81). Both classes are well recognized, with only 370 misclassifications of class E, and 350 misclassifications of class I. The disproportion between the number of introverts and extroverts is shown in the set with a large predominance of introverts (approx. 77% of the set). The relatively small number of incorrect assignments indicates that the model effectively distinguished introverts and extroverts based on their written test.

Table 8: Confusion matrix – gemini-2.0-flash – letter I/E

	Predicition I	Predicition E

True I	6034	350
True E	370	1540

As shown in Table 9 there is strong asymmetry in recognition of letter N/S by gpt-3.5-turbo model. The model exhibits strong bias toward class N, misclassifying over half of actual class S examples (approx. 56%). In case of this letter is also shown big disproportions between the number of intuition and sensing types with large predominance of intuition types (approx. 86% of the set). This results suggest the gpt-3.5-turbo model struggles with correct recognition of sensing type.

Table 9: Confusion matrix – gpt-3.5-turbo – letter N/S

	Predicition N	**Predicition S**
True N	7089	65
True S	641	499

The confusion matrix in Table 10 is the example of well-balanced recognition of F/T letter by gpt-4o-mini model. Of the 4491 true class feeling (F) 4090 were correctly recognized, while 401 were misclassified. In class thinking (T), 3253 out of the 3803 were correctly recognized, while 550 were misclassified. This distribution reflects the high F1-scores for both classes and suggests reliable distinction between the feeling and thinking traits.

Table 10: Confusion matrix – gpt-4o-mini – letter F/T

	Predicition F	**Predicition T**
True F	4090	401
True T	550	3253

According to Table 11, the confusion matrix for letter J/P for gemini-2.0-flash reviles an assymetry in classification. Of the 3307 true class judging (J) 2140 were correctly recognized, while 1167 were misclassified. In class perceiving (P), 4871 out of the 4987 were correctly recognized, while 116 were misclassified. This results are suggesting that the model exhibits overclassification of the overrepresented perceiving class (approx. 60% of records).

Table 11: Confusion matrix – gemini-2.0-flash – letter J/P

	Predicition J	**Predicition P**
True J	2140	1167
True P	116	4871

The last important result of this experiment is exact-match accuracy over the full 4-letter string. As shown on Figure 2, the best predictions were delivered by gemini-2.0-flash model (0.72), then a slightly lower performance was shown by gpt-4o-mini (0.7) and the lowest of the tested model with visibly weaker performance was gpt-3.5-turbo (0.63).

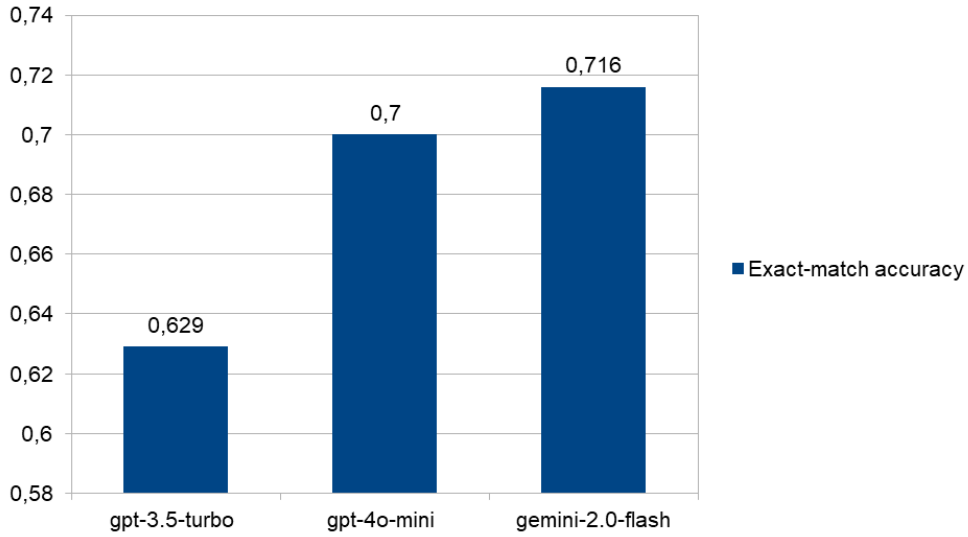

Figure 2: Exact-match accuracy over the full 4-letter string

Conclusions

This study aimed to evaluate and compare performance of three large language models (LLMs): gpt-3.5-turbo, gpt-4o-mini and gemini-2.0-flash in the task of MBTI personality classification based on posts. We focused on each seperate letter prediction as well as full four letter string classification. The results analysis shown notable differences between models and in classification of single letters. Across the evaluated models gemini-2.0-flash achieved the highest exact-match accuracy and had the best balance of accuracy and recall for all MBTI letters. The best performance was achieved by the model in recognizing emotional-logical (F/T) and judging-perceiving (J/P) traits. The gpt-3.5-turbo exhibited strong accuracy but in a few letter prediction demonstrated tendency to misclassification of minority class. This is especially visible on letter N/S in which case model achieved reasonably high accuracy, showing that accuracy can be misleading under class imbalance. Although, the gpt-4o-mini demonstrated good balance of metrics and shows visible improvement over the gpt-3.5-turbo it does not exceed the performance of the gemini-2.0-flash. The model exhibited high performance a specially on letters T/F and J/P with high accuracy in recognition of perceiving (P) trait but lower recall for the judging class (J) introducing asymmetry in class recognition.

Among all evaluated models, the results demonstrated that high accuracy may conceal weak performance on minority classes, particularly in imbalanced datasets. Furthermore, letter-specific consistent tendencies were observed: letter N/S was the most demanding (especially for gpt-3.5-turbo model), on the other hand, models demonstrated higher reliability in predicting letters F/T and J/P.The I/E letter also yielded strong results, though it exhibited higher divergence in predictions among the tested models. This analysis illustrates the critical role of using class-sensitive metrics like recall and F1-score in evaluating multi-label classification tasks, especially those involving subtle human characteristics. Notably, this research relied on publicly available personality-labeled datasets, which, in addition to being restricted to MBTI codes, may contain subjective or inconsistent annotations. These outcomes imply that while LLMs are capable of extracting underlying personality traits from textual data, their reliability diminishes in the presence of imbalanced class distributions. Future work should investigate the impact of class balancing, fine-tuning to ensure fairer and more accurate predictions. Additionally, extending the evaluation to alternative personality frameworks – such as the Big Five – exploring LLM performance across languages and text types, as well as addressing the interpretability and ethical dimensions of personality classification, would provide meaningful directions for further study.

Acknowledgment

This work was supported by the Military University of Technology under Project UGB 531-000023-W500-22.

References

- Eva Eigner and Thorsten Händler, ‘Determinants of LLM-assisted Decision-Making’, arXiv, 2024, doi: 10.48550/arXiv.2402.17385.
- Song Lai, Bo Sun, Fati Wu, and Rong Xiao, ‘Automatic Personality Identification Using Students’ Online Learning Behavior’, IEEE Trans. Learn. Technol., vol. 13, no. 1, 2020.
- Leong Dickmond, Vazeerudeen Abdul Hameed, and Muhammad Ehsan Rana, ‘A Study of Machine Learning Based Approaches to Extract Personality Information from Curriculum Vitae’, presented at the 2021 14th International Conference on Developments in eSystems Engineering (DeSE), IEEE, 2021. doi: 10.1109/DESE54285.2021.9719496.
- Lakshyajit Thapa, Aniket Pandey, Deepanshu Gupta, Ayush Deep, and Rakesh Garg, ‘A Framework for Personality Prediction for E-Recruitment Using Machine Learning Algorithms’, presented at the 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2024. doi: 10.1109/CONFLUENCE60223.2024.10463354.
- J. Cui et al., ‘Machine Mindset: An MBTI Exploration of Large Language Models’, Jun. 02, 2024, arXiv: arXiv:2312.12999. doi: 10.48550/arXiv.2312.12999.
- Isabel Briggs Myers, Mary H. McCaulley, Naomi L. Quenk, and Allen L. Hammer, MBTI Manual A Guide To The Development And Use Of The Myers Briggs Type Indicator. Paolo Alto, California: CONSULTING PSYCHOLOGY PRESS, INC, 1998.
- C. Basto, ‘Extending the Abstraction of Personality Types based on MBTI with Machine Learning and Natural Language Processing’, May 25, 2021, arXiv: arXiv:2105.11798. doi: 10.48550/arXiv.2105.11798.
- V. G. dos Santos and I. Paraboni, ‘Myers-Briggs personality classification from social media text using pre-trained language models’, arXiv.org. Accessed: May 04, 2025. [Online]. Available: https://arxiv.org/abs/2207.04476v1
- Dr. Curry Guinn, ‘Assessing Author Personality Types Using ChatGPT’, presented at the 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE), IEEE, 2023. doi: 10.1109/CSCE60160.2023.00021.
- E. Wang and H. Wang, ‘MBTI Personality Recognition and Performance Improvement in LLMs’, Feb. 06, 2025, Social Science Research Network, Rochester, NY: 5111274. doi: 10.2139/ssrn.5111274.
- J., Mitchell, ‘(MBTI) Myers-Briggs Personality Type Dataset’. kaggle.com, https://www.kaggle.com/datasets/datasnaek/mbti-type, 2017.