applied sciences

MDPI

Article

Synthetic Text as Data: On Usefulness and Limitations

Seoyeon Choi [1,†], Jaein Sim [2,†] and Guebin Choi [2,*]

[1] Research and Development Center, Nanum Space Co., Ltd., Jeonju 54907, Republic of Korea; chltjdus1212@gmail.com
[2] Department of Statistics, Institute of Applied Statistics, Jeonbuk National University, Jeonju 54896, Republic of Korea; jisim@jbnu.ac.kr
* Correspondence: guebin@jbnu.ac.kr; Tel.: +82-63-270-3387
† These authors contributed equally to this work.

Abstract: This study investigates the utility of GPT-generated text as a training resource in supervised learning, focusing on two perspectives: its effectiveness as an augmentation tool in data-scarce or class-imbalanced settings and its potential as a substitute for human-written data. Using MBTI personality classification as a benchmark task, we conducted controlled experiments under both class imbalance and few-shot learning conditions. Results showed that GPT-generated text could improve classification performance when used to supplement underrepresented classes. However, when synthetic data fully replace real data, performance declines significantly—particularly in tasks requiring fine-grained semantic distinctions. Further analysis reveals that GPT outputs often capture only partial personality traits, enabling coarse-level classification but falling short in nuanced cases. These findings suggest that GPT-generated text can function as a conditional training resource, with its effectiveness closely tied to the granularity of the classification task.

Keywords: GPT-generated text; synthetic data; data augmentation; MBTI classification; class imbalance; few-shot learning; fine-grained classification; data granularity; large language models

check for updates

Academic Editor: Christos Bouras

Received: 15 April 2025
Revised: 4 May 2025
Accepted: 12 May 2025
Published: 13 May 2025

Citation: Choi, S.; Sim, J.; Choi, G. Synthetic Text as Data: On Usefulness and Limitations. *Appl. Sci.* **2025**, *15*, 5460. https://doi.org/10.3390/app15105460

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Large language models (LLMs), such as ChatGPT, have demonstrated remarkable capabilities in generating fluent, contextually appropriate, and human-like text. This has opened new possibilities in natural language processing—not only in text generation tasks but also in augmenting datasets for downstream applications. However, a fundamental question remains underexplored: can LLM-generated text be treated as reliable training data?

To be more specific, if we replace or supplement real-world data with synthetic text generated by GPT, how close will the resulting model performance be? Will it be identical, noticeably worse, or somewhere in between? And, if it is somewhere in between, can we quantify the gap? These are critical questions for understanding whether GPT-generated text is merely stylistically convincing or genuinely informative in the way real data are. Addressing this issue requires moving beyond the generation quality and evaluating synthetic text from a data-centric perspective.

In this study, we investigate the representational value of GPT-generated text by comparing its utility to that of real data in supervised learning. Our working hypothesis is straightforward: if GPT-generated samples are as informative as real samples, models trained on them should perform similarly. If there is a significant performance gap, then despite the surface-level fluency, synthetic text may lack task-relevant information present

in human-written examples. In this sense, we view classification accuracy not just as a metric of model performance but as an indirect proxy for data quality.

To construct a measurable evaluation framework, we use MBTI personality classification as a benchmark task. This dataset features both class imbalance and low-resource conditions, which we deliberately leverage—not as problems to solve, but as controlled scenarios in which the effectiveness of synthetic data becomes testable. Experiment 1 evaluates whether GPT-generated text can mitigate class imbalance, while Experiment 2 simulates a few-shot setting to assess how well synthetic samples substitute for scarce real data. In both cases, we compare models trained with real, synthetic, and mixed data to measure the relative utility of each.

Our findings lead to two main contributions. First, we show that GPT-generated samples can supplement real training data in meaningful ways. In settings involving data scarcity—such as class imbalance or few-shot learning—synthetic augmentation leads to measurable improvements in model performance. This suggests that GPT can serve as a practical augmentation tool when real data are limited. Second, we also clarify the limitations of synthetic text. While useful as a supplement, it does not function as a full substitute: models trained solely on GPT-generated data consistently underperform, particularly in fine-grained classification tasks. This emphasizes the importance of treating GPT-generated content not as a direct replacement for real data but as a conditional support whose value depends heavily on task complexity and granularity. Our study thus contributes both a practical takeaway and a cautionary insight into the role of LLMs in data-centric AI workflows.

2. Related Works

This study centers on two main questions: First, can text generated by large language models (LLMs) such as GPT be used effectively for data augmentation? Second, can such synthetic text function as a valid substitute or supplement for real-world training data? Accordingly, we review related works along these two lines of inquiry.

We begin with prior research on data augmentation. In natural language processing (NLP), various augmentation techniques have long been used to address class imbalance and low-resource conditions. Traditional methods such as synonym replacement, back-translation, and random insertion or deletion offer computational simplicity and ease of implementation [1,2]. However, these methods often fall short in preserving semantic coherence or contextual appropriateness [3]. With the recent advancements in LLMs, especially the GPT series, a new line of work has emerged focusing on generating more fluent and context-aware synthetic text for data augmentation. For instance, Dai et al. [4] rephrased training samples using ChatGPT to improve classification accuracy. Abaskohi et al. [5] leveraged prompt-based paraphrasing with contrastive learning for effective few-shot classification. Oh et al. [6] applied GPT-generated text to improve performance in neural machine translation, and Cegin et al. [7] demonstrated that ChatGPT-based paraphrases show comparable diversity and robustness to human-written examples in intent classification tasks. While these approaches have mainly been applied to binary or sentence-level tasks, there is growing interest in expanding their application to more fine-grained and complex classification problems, where their potential remains to be further explored.

We now turn to the question of synthetic text validity. While GPT-generated text is often fluent and grammatically correct—sometimes indistinguishable from human-authored content on the surface—its usefulness as training data is still under investigation. However, the question of whether GPT-generated text is truly useful as training data remains actively debated. Prior studies have primarily focused on leveraging the generative abilities of LLMs to perform specific tasks [8,9], while relatively few have quantitatively

evaluated the effectiveness of synthetic text when used directly as training input. In response, a number of recent works have begun to examine the informativeness and generalizability of LLM-generated content. For instance, Zhang et al. [10] found that models trained on data generated by other LLMs tend to underperform compared to those trained on real data. McCoy et al. [11] observed that GPT-generated samples often repeat known patterns and lack novelty, while Dou et al. [12] proposed the Scarecrow framework to systematically analyze errors in GPT-3 outputs and highlight subtle distinctions from human-authored text. Li et al. [13] systematically explored this issue by evaluating the performance of classifiers trained on GPT-generated text across various tasks. Their results show that while synthetic data can be valuable in low-subjectivity settings, its effectiveness significantly declines in more subjective or fine-grained tasks.

Building on these discussions, our study aims to evaluate the value of GPT-generated text as a training resource, either as a supplement to or substitute for real data. In particular, we investigate how such synthetic samples perform under data-scarce and imbalanced conditions, offering a quantitative analysis of their practical utility across different experimental settings.

3. Data Explanation

3.1. Kaggle Data

MBTI (Myers–Briggs Type Indicator) is a personality assessment based on Carl Jung's theory of psychological types. It classifies individuals into 16 personality types by combining four dichotomous dimensions: extraversion–introversion, sensing–intuition, thinking–feeling, and judging–perceiving. The dataset used in this study was obtained from a publicly available MBTI dataset on Kaggle [14], which compiles forum posts from users of PersonalityCafe, an online community. Each sample consists of the user's MBTI label and their recent posts. The text is informal and conversational in nature, often including abbreviations, slang, and internet-specific language, presenting unique challenges for natural language processing (NLP) tasks. This dataset includes 8675 user samples, each consisting of an MBTI label—one of the 16 types—and the user's 50 most recent forum posts on PersonalityCafe. A summary of the dataset is provided in Table 1.

Table 1. Sample text posts classified by MBTI types in [14].

Type	Posts
ENTP	I'm finding the lack of me in these posts ...
INTJ	Dear INTP, I enjoyed our conversation the ...
INTJ	I tend to build up a collection of things ...

This dataset is particularly suitable for studying text data augmentation for several reasons. First, it presents a severe class imbalance: some types such as INFP and INTJ comprise over 1000 samples each, whereas others like ENFJ and ESFP appear fewer than 100 times. Figure 1 illustrates the distribution of sample counts across the 16 MBTI personality types in the dataset.

This imbalance reflects a common real-world challenge in natural language datasets and offers a realistic setting to test whether synthetic data can improve minority class representation. For example, while the INFP class contains 1832 entries, ESTJ appears only 39 times—making this a practical scenario for evaluating the impact of augmentation. First, a manual inspection was conducted as a minimal verification step to ensure that the generated texts were appropriate for their corresponding personality types. Second, the dataset naturally gives rise to a few-shot learning situation for the underrepresented types.

In conventional supervised learning, classes with only a handful of examples typically suffer from poor generalization. Augmenting those classes with GPT-generated text provides a direct way to assess how language models can alleviate data scarcity. Third, the dataset is composed of informal, user-generated text spanning personal anecdotes, reflections, and opinions. This variability in tone, vocabulary, and topic makes it a challenging yet realistic test case for generating stylistically consistent synthetic samples using LLMs.

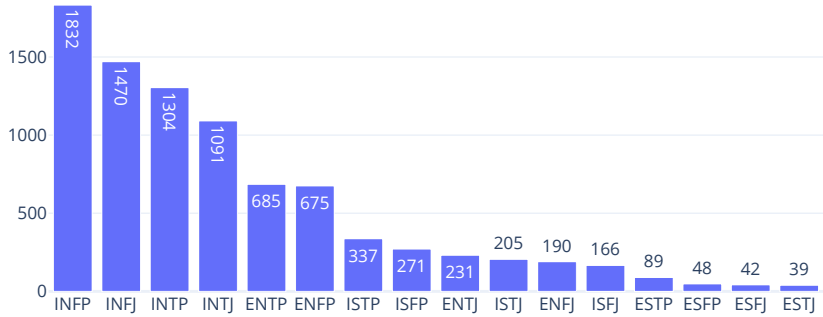


Figure 1. The distribution of sample counts across the 16 MBTI personality types in the dataset.

In summary, the MBTI dataset serves as an ideal case study for this work, as it simultaneously exhibits class imbalance, few-shot conditions, and stylistic diversity—three key factors that motivate the need for high-quality text augmentation.

3.2. GPT-Generated Data

To address the class imbalance present in the MBTI dataset, we generated additional text samples using OpenAI's GPT-3.5 model via the ChatGPT interface. While the primary objective was to supplement underrepresented types, this process also offered a unique opportunity to examine whether GPT-generated text could approximate the value of human-written data in training machine learning models. The generation process involved a carefully designed prompt template for each of the 16 MBTI types. Each prompt instructed GPT to write a scenario-based Instagram post that reflects the personality traits and linguistic style of a given MBTI type. For example, the prompt for ESTJ was the following:

> *"Could you please write a scenario for an Instagram post that an ESTJ might share, with various situations, totaling 1500 characters? Write it in a way that resembles an Instagram caption, and exclude emojis and hashtags."*

The same structure was applied to all types by substituting the MBTI label. While the prompt design was uniform, maintaining stylistic consistency and personality-specific nuance proved challenging. GPT sometimes produced overly generic or repetitive outputs, and manual inspection was needed to ensure quality and diversity. To rebalance the dataset, we generated 200 samples for each low-resource type (e.g., ESTP, ESFP, ENFJ), aligning the output length with the original dataset by capping responses at 1500 characters.

The generation process involved several practical challenges. First, a minimal level of manual inspection was required to ensure the validity of the generated text. Second, label leakage was a recurring issue—some responses included explicit type references (e.g., "As an ENFJ, I often ..."), which were removed during post-processing to avoid biasing the classifier. Third, the outputs showed significant variability in length. Posts under 300 characters were often too sparse to be informative, while those exceeding 1800 characters added unnecessary training cost. We applied length-based filtering to ensure consistency. Details for these points are available in the Appendix A. Additionally, although the prompts explicitly prohibited emojis and hashtags, some appeared and were manually removed during cleanup.

Balancing automation and manual oversight was crucial. While GPT enabled scalable and stylistically fluent generation, ensuring semantic variety and task relevance required iterative refinement. Ultimately, the synthetic samples were combined with the original dataset and used to train MBTI classification models.

Importantly, these synthetic texts served two distinct purposes in our study. First, they acted as a form of targeted augmentation to improve model performance in imbalanced or low-resource settings. Second, and more critically, they enabled us to investigate a broader question: whether LLM-generated content can function as a valid substitute—not merely a supplement—for human-authored data. This dual role reflects the central focus of our study and frames how the results are interpreted in later sections.

4. Experiments

4.1. Experiment Overview

This study evaluates the value of GPT-generated text as a training resource from two distinct perspectives. The first is whether such text can serve as an effective augmentation tool in data-scarce or class-imbalanced settings. The second is whether it can match the informativeness of human-written text and serve as a potential substitute. Each experiment in this paper was designed to address one of these two questions.

The experiment described in Section 4.2 focused on the first question. While the experimental setup was framed as a class imbalance problem, the underlying goal was to assess whether GPT-generated text contributed meaningfully to model learning when used to supplement underrepresented classes. In this sense, the class imbalance condition served as a testbed for evaluating the utility of synthetic data as an augmentation resource.

In contrast, the experiment described in Section 4.3 was designed to address the second question: can GPT-generated text approximate the functional role of real data? To test this, we adopted a few-shot learning setup. Compared to class imbalance, few-shot conditions provided a more stringent and focused environment in which all classes had very limited data. This made it a more appropriate setting for examining whether synthetic data can serve as a viable alternative to genuine samples.

4.2. Evaluating GPT-Generated Text as a Data Augmentation Tool

The first experiment examines whether GPT-generated text can serve as an effective data augmentation tool for addressing class imbalance in text classification. Although the experimental structure was framed around an imbalanced dataset, the underlying goal was to assess the utility of synthetic data in improving model performance for underrepresented classes. This corresponds to the first of our two research questions: can GPT-generated text function as a meaningful supplement to real data?

The experiment was conducted under two settings: the Base Dataset and the Extended Dataset. In the Base Dataset setting, we used the original MBTI dataset from Kaggle [14] and split each MBTI type equally into 50% training and 50% test sets. Although a 1:1 train–test split is not standard in typical machine learning practice, we deliberately adopted this ratio to ensure that even low-resource classes would have enough test samples to allow stable and interpretable evaluation. This decision aligned with the core objective of the study—not to build the most optimized model, but to assess the contribution of GPT-generated text under controlled and comparable conditions.

The Extended Dataset setting built on the Base Dataset by supplementing the training data for four low-resource MBTI types—ESTP, ESFP, ESFJ, and ESTJ—with synthetic samples generated by GPT. For each of these types, we increased the number of training examples to a total of 200 by combining the original Kaggle data with GPT-generated posts, while keeping the test sets unchanged. In this setup, the Base Dataset consisted of

a straightforward 50:50 split of the original data, whereas the Extended Dataset added synthetic text to reach 200 training examples per class for the selected types. The resulting sample distribution is visualized in Figure 2.

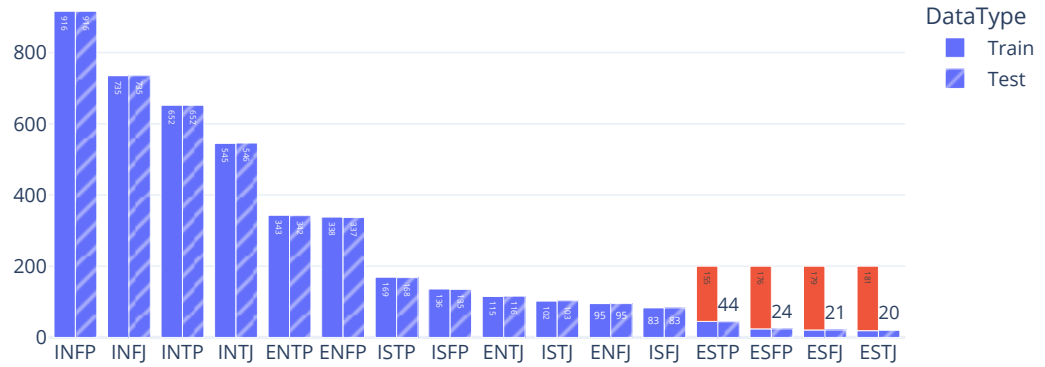


Figure 2. The sample distribution for the Extended Dataset scenario. Solid bars represent training samples and hatched bars indicate test samples for each MBTI type. Additional training data generated by GPT are shown in red for the last four types (ESTP, ESFP, ESFJ, ESTJ).

This design allowed us to directly assess whether GPT-generated text contributed to model improvement. If performance improved, we could infer that synthetic text was useful in supplementing scarce data. Conversely, if no improvement was observed, the setup still offered clear evidence that the synthetic data may have lacked sufficient informativeness or compatibility with the learning model.

Our analysis followed a conventional approach to text classification, combining traditional feature-based methods with machine learning models. We began by cleaning and tokenizing the text data, removing stop words and special characters, and applying basic normalization. From the preprocessed text, we generated 3-g features—sequences of three consecutive tokens. Each document was then represented as a frequency vector based on these N-grams and used as input to the classification models. For classification, we experimented with a variety of models, including gradient boosting methods such as LightGBM [15,16], CatBoost [17], and XGBoost [18]; ensemble-based methods such as Random Forest and Extra Trees; a distance-based method (K-Nearest Neighbors); a simple feedforward neural network; and a weighted ensemble meta-model [19] that combined the outputs of individual classifiers. Model performance was evaluated on the test set using precision, recall, F1 score, and Area Under the Curve (AUC).

The experimental results for this study are available at https://seoyeonc.github.io/MBTI_dashboard (accessed on 10 May 2025), with the results for the experiment in the current section presented under Experiment 1 on the dashboard. The dashboard was built using the Quarto publishing system [20]. Simulations were conducted using `Autogluon` [21,22], which enabled automated model selection and evaluation. Due to the large volume of results, the main text presents only a selected subset of findings that are most relevant for discussion. We limited our analysis to the top six models based on overall classification performance. Among various evaluation metrics, we report only the AUC, as it offers a threshold-independent and stable measure of classification performance—particularly suitable for comparing models in imbalanced data settings.

Figure 3 compares the AUC scores of six top-performing models across four underrepresented MBTI types (ESFP, ESTP, ESTJ, ESFJ). The blue bars represent performance using only the original Kaggle dataset, while the red bars reflect results from the Extended Dataset, which included GPT-augmented samples. Across most combinations of model and personality type, the inclusion of GPT-generated text led to performance improvements. For instance, models such as LightGBM and CatBoost showed notable gains for ESTP and

ESFP when augmented data were added. Although the magnitude of improvement varied by model, the overall trend remained consistent: synthetic data contributed positively to classification performance. In no case did performance significantly deteriorate due to augmentation, and in several cases, it helped close the gap caused by data scarcity. Taken together, the results suggest that although the degree of improvement varied across models and MBTI types, GPT-based augmentation consistently led to performance gains. While there were a few exceptions where augmentation did not yield clear improvements (as shown for XGBoost and some other models in the dashboard, possibly due to differences in how the model processes textual features or responds to synthetic inputs), the majority of cases showed that models trained on augmented datasets performed better than those trained on the original data alone. In fact, for every MBTI class examined, the highest AUC score was always achieved using the augmented dataset—a strong indication that synthetic data are not only helpful but at times critical in overcoming data scarcity. These results provide empirical support for our first research question—whether large language model outputs can be leveraged as an effective data augmentation strategy. Having established the utility of GPT-augmented data as a supplemental resource, we now turn to a deeper question: can synthetic text also function as a substitute for real-world data? In other words, if we had replaced the GPT-generated samples with actual human-written text, would the performance have improved even more? To address this, we designed a second experiment under a few-shot learning setup, which provided a more extreme and controlled scenario for testing such substitution. We describe this next.

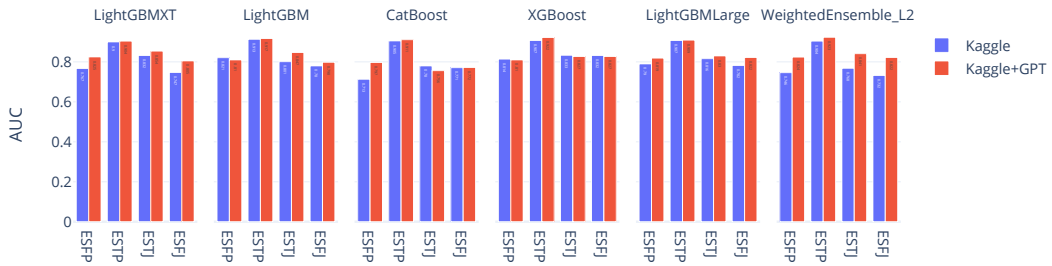


Figure 3. The bar chart compares the AUC scores of various machine learning models across additional MBTI classes, with results based on datasets exclusively from Kaggle (blue) and those augmented with GPT-generated data (red).

4.3. Assessing the Substitutive Value of GPT-Generated Text

This section investigates the extent to which GPT-generated text can serve as a substitute for real data. In particular, we ask whether such text goes beyond mere surface-level fluency and stylistic plausibility to contain information that is genuinely useful for a classification task. The central question is whether GPT-generated text can function not just as a supplemental augmentation resource but as a training signal of equivalent value to human-written data. This distinction can be illustrated with an analogy from image processing. Augmented images—created via transformations such as rotation, cropping, or color adjustment—are ultimately derivatives of existing data. In contrast, a new real-world image captures independent information from a novel viewpoint. As such, the latter typically provides more diverse and informative content, which can contribute more significantly to model learning. The aim of this section is to determine whether GPT-generated text belongs in the former or the latter category—whether it can match the informational value of truly independent samples.

To test this, we constructed a few-shot learning scenario. By limiting the amount of data available for all classes, we were able to observe how heavily the model relied on the training samples and thereby assess the informational depth of GPT text. The experiment centered

on ENFJ, a relatively underrepresented MBTI type in the Kaggle dataset. The training set included only a single real ENFJ example, while the test set contained 30 fixed ENFJ examples. Under this setup, we compared two data addition conditions: one in which the training set was augmented with GPT-generated ENFJ text and the other in which it was augmented with the same number of real ENFJ samples from the remaining Kaggle data. The structure and distribution of the training and test sets are illustrated in Figure 4. In this figure, hatched bars indicate test samples, and red bars represent additional GPT-generated samples added to the training set. Blue bars denote real data from the original Kaggle corpus.

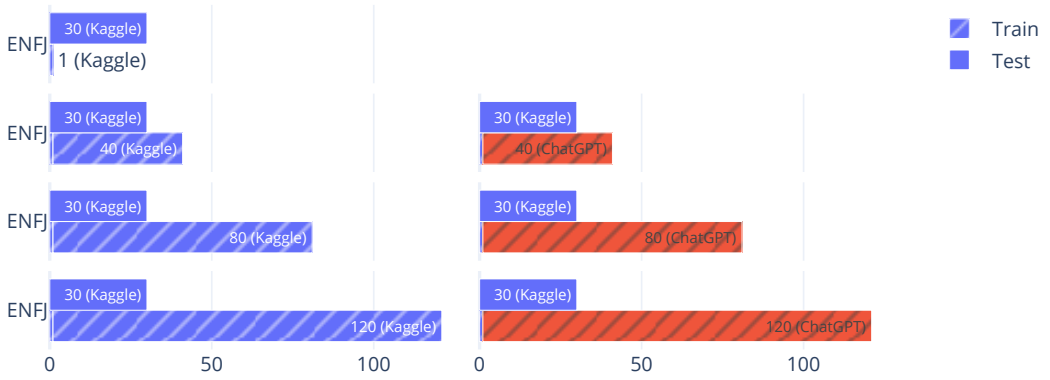


Figure 4. Visualization of ENFJ data distribution in training and testing sets, comparing the baseline Kaggle dataset against incremental GPT-augmented data integration. The *x*-axis quantifies the number of samples. Blue bars signify samples derived solely from Kaggle data, visible on the left side of the chart, and red bars represent additional GPT-generated samples included for the training set, shown on the right side.

The experimental scenario was designed based on the following ideas. First, we artificially reduced the number of training samples for a specific class to simulate an extreme few-shot learning condition. Under such circumstances, the model's performance was inevitably limited, and it was reasonable to expect that adding more training samples would lead to improved performance. Second, we established two experimental conditions: one in which only GPT-generated synthetic data were added for the target class and another in which the same number of real samples were added instead. As a result, one experiment involved a training set for the target class composed almost entirely of synthetic data, while the other used only real data. If the performance improvement achieved by adding synthetic data was substantially lower than that achieved by adding real data, it would be difficult to consider synthetic data equivalent to real data in providing an effective learning signal. This experiment differed significantly from the previous experiment. In the earlier setting, both synthetic and real data were included across all classes, and the role of synthetic data was evaluated primarily as a supplemental augmentation resource. In contrast, the current experiment isolated the effect by including only synthetic or only real data for a specific class, thereby directly assessing whether synthetic data can function as an equivalent training signal to real data.

The experimental results are summarized in the Experiment 2 tab at https://seoyeonc.github.io/MBTI_dashboard (accessed on 10 May 2025), with only a subset visualized in Figure 5. Overall, we found that GPT-generated text could improve performance relative to the one-shot baseline. However, its effectiveness remained consistently below that of real data, even when the quantity of synthetic samples increased. This suggests that while GPT outputs may appear fluent and coherent on the surface, they often lack the subtle, class-discriminative signals necessary for fine-grained classification.

In conclusion, GPT-generated text appears to be a useful supplement to training data, especially under conditions of scarcity. However, it falls short of functioning as a true substitute for real-world samples in terms of informational richness and diversity. This finding highlights a key limitation and provides an important cautionary insight into the role of synthetic text in supervised learning tasks.

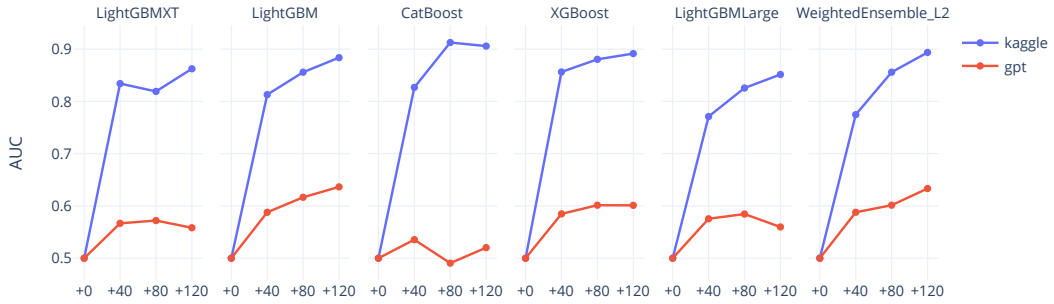


Figure 5. The graph illustrates the AUC results for the classification of ENFJ versus other MBTI types, with the *x*-axis indicating the incremental addition of data (+0, +40, +80, +120 samples), and the *y*-axis representing the AUC values. The blue line shows the results when actual data samples were added, while the red line corresponds to the results with the inclusion of GPT-generated synthetic data.

The previous experiment demonstrated a consistent performance gap between GPT-generated text and real-world data in fine-grained classification. However, while this result was clear, it also raised a new question: does such a difference imply a binary distinction between usefulness and uselessness, or might synthetic data still retain partial informativeness? Rather than concluding that GPT-generated text is entirely lacking in value, we sought to explore whether it might preserve some—but not all—of the characteristics necessary for accurate classification.

For instance, consider the ENFJ personality type. To convincingly replicate an ENFJ-written post, a model must reflect all four dimensions—E, N, F, and J. If GPT-generated text reflects only some traits (e.g., E and N), it may not resemble a complete ENFP profile, but rather a partial form like EN−− or E−−J (in this notation, the "−" symbol indicates an MBTI trait that is unclear or not strongly represented in the text). In such cases, the synthetic text may enable coarse-grained differentiation (e.g., extroverted vs. introverted) but fail in more nuanced distinctions between similar personality types. This motivated a refined hypothesis: while GPT-generated content may underperform in subtle classifications, it could still be effective when distinguishing between types that are more linguistically and conceptually distant.

To examine this hypothesis, we designed a binary classification task in which ENFJ is contrasted against five MBTI types presumed to be relatively dissimilar: ISTP, INTP, ISFP, ISTJ, and ESTP. The experimental configuration followed the same few-shot structure as before. The ENFJ class contained only a single real training example, with 30 fixed test samples. GPT-generated ENFJ texts were added to the training set in increments of 20, up to a total of 140 samples. The data distributions for the comparison types were kept consistent with the original Kaggle dataset.

The full results of this experiment are available in the Experiment 3 tab at https://seoyeonc.github.io/MBTI_dashboard (accessed on 10 May 2025), with key findings visualized in Figure 6. Interestingly, the results diverged from previous patterns. For several classifiers, performance improved steadily as more GPT-generated data were added, and in some cases, the models approached AUC levels comparable to those of models trained on real data. While overall performance still fell short of the real-data baseline, the synthetic ENFJ text was nevertheless sufficient to distinguish ENFJ from markedly different personality types with reasonable accuracy.

These findings suggest that GPT-generated text is not universally ineffective. In classification settings where the decision boundary is coarse and clearly defined, such as separating ENFJ from conceptually distant types, synthetic data can still serve a functional role. In contrast, more fine-grained tasks—where subtle semantic or stylistic differences are required—appear to exceed the representational capacity of GPT-generated content.

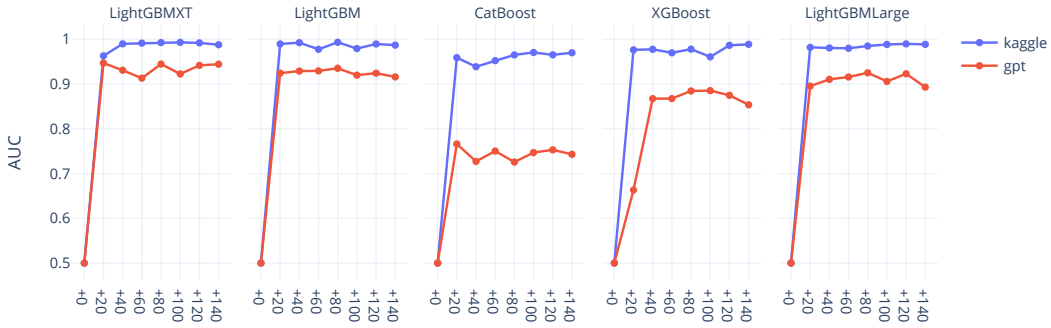


Figure 6. This chart depicts the AUC metrics for differentiating ENFJ from various MBTI categories, with steps of data augmentation displayed on the x-axis (+0, +40, +80, +120 samples), and the corresponding AUC performance on the y-axis. The outcomes when incorporating real-world data samples are traced by the blue line, while the red line illustrates the effects of including GPT-generated data into the analysis

Taken together, this experiment highlights a critical insight: the limitations of synthetic data are not absolute but conditional. GPT-generated samples may behave similarly to real data in broad categories while diverging in finer details. In this sense, the distinction lies not solely in accuracy metrics, but in the grain of information captured. This distinction calls for a more nuanced understanding of generative data—one that accounts for the structure, depth, and specificity of the information being modeled.

5. Conclusions and Discussion

This study set out to evaluate the utility of GPT-generated text as training data for supervised learning tasks. We approached this question from two complementary angles: first, whether synthetic text can effectively supplement real data in settings of class imbalance or data scarcity; and second, whether it can substitute for human-written text by providing comparable informational depth.

Using MBTI personality classification as a benchmark task, we conducted a series of experiments under controlled low-resource conditions. Our findings indicate that GPT-generated text can serve as a practical augmentation resource—improving model performance when added to underrepresented classes. In particular, models trained on mixed datasets (real + synthetic) consistently outperformed those trained on real data alone in imbalanced settings. These results affirm the value of GPT outputs as a cost-effective means of extending datasets in situations where collecting real samples is difficult.

However, our analysis also reveals a clear gap between synthetic and real data when it comes to substitutive value. In few-shot learning experiments, models trained solely on GPT-generated text underperformed those trained on equivalent real data, even when the amount of synthetic data was substantially increased. This suggests that surface-level fluency does not guarantee task-relevant informativeness. Further investigations showed that while GPT-generated text may have failed to capture the full richness of real samples, it retained partial signal strength—enough to distinguish between personality types that are more dissimilar, but insufficient for subtle, fine-grained classification tasks.

These findings point to an important distinction in how synthetic data should be interpreted and deployed. Rather than viewing GPT-generated text as either inherently

valid or entirely inadequate, we argue that its effectiveness is highly dependent on the granularity of the classification task. Coarse distinctions may be well supported by generative models, whereas nuanced tasks likely demand the semantic depth found only in real human-authored data. In conclusion, GPT-generated text shows promise as a conditional training resource—useful in scenarios with limited real data, particularly when the classification target is broad or well-separated.

Future research should explore the generalizability of the findings to other NLP tasks, the linguistic characteristics of synthetic text, comparisons with newer language models, and strategies for mitigating potential biases and fairness risks. While MBTI classification serves as a convenient benchmark, additional validation is needed to determine whether the results generalize to other NLP tasks such as sentiment analysis or intent classification. Moreover, this study did not conduct an in-depth analysis of the linguistic properties of synthetic text, nor did it include comparisons between GPT-3.5 and newer models such as GPT-4o. The observed variation in performance across personality types and models also warrants systematic investigation. Finally, exploring approaches to mitigate the potential biases and fairness risks introduced by synthetic data (e.g., as discussed in [23]) remains an important direction for future work.

Author Contributions: Conceptualization, G.C.; Methodology, S.C.; Formal analysis, S.C.; Data curation, J.S.; Writing—original draft, S.C. and J.S.; Writing—review & editing, G.C.; Visualization, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Research Foundation of Korea (NRF) funded by the Korea government (RS-2023-00249743).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All code used in this study is available at https://github.com/seoyeonc/TextSyn (accessed on 10 May 2025).

Conflicts of Interest: Author Seoyeon Choi was employed by the company Nanum Space Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

Appendix A.1. First Challenge of Answers of ChatGPT

To verify whether each text matched the intended category, we conducted manual inspection. For example, in Figure A1, we checked whether the response aligned with the extroversion (E) category by examining the red-underlined expressions. Figures A2–A8 were presented to illustrate this manual checking process (for illustrative purposes only; not used in model training).

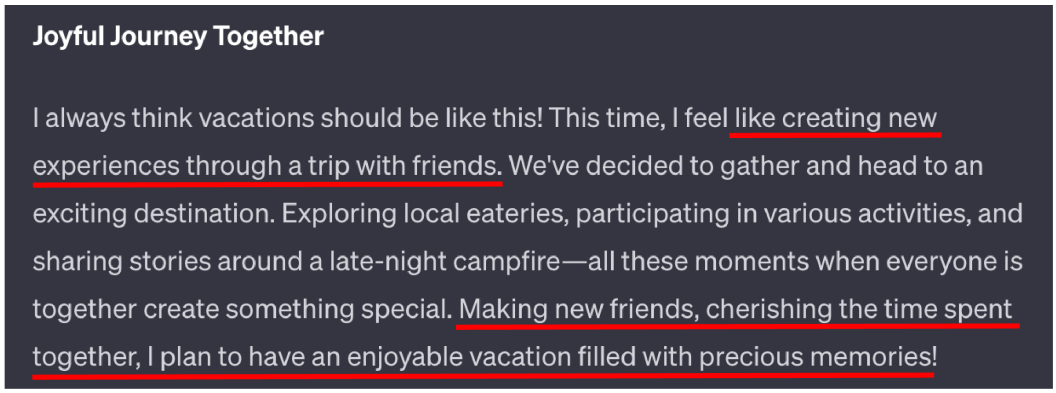


Figure A1. An answer in the category of extroversion (E) from ChatGPT. Red underlines indicate expressions that reflect extroverted tendencies, such as social engagement, shared experiences, and emotional emphasis.

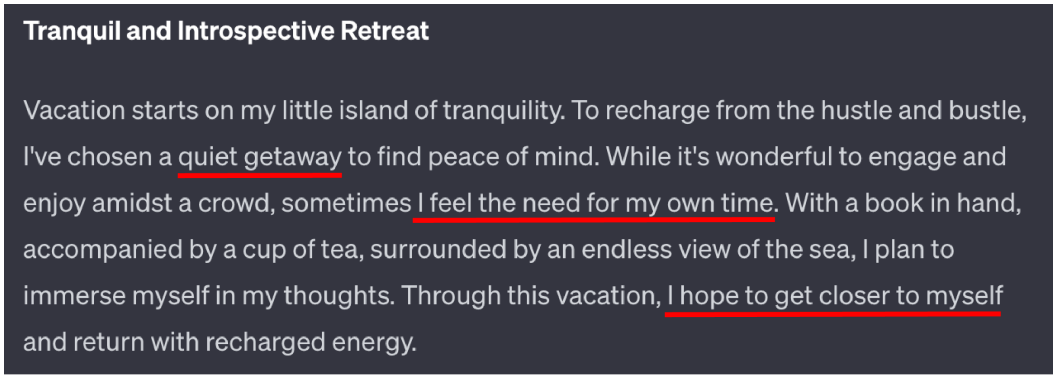


Figure A2. An answer in the category of introversion (I) from ChatGPT. Red underlines indicate expressions that reflect introverted tendencies, such as a preference for solitude, self-reflection, and personal emotional recharge.

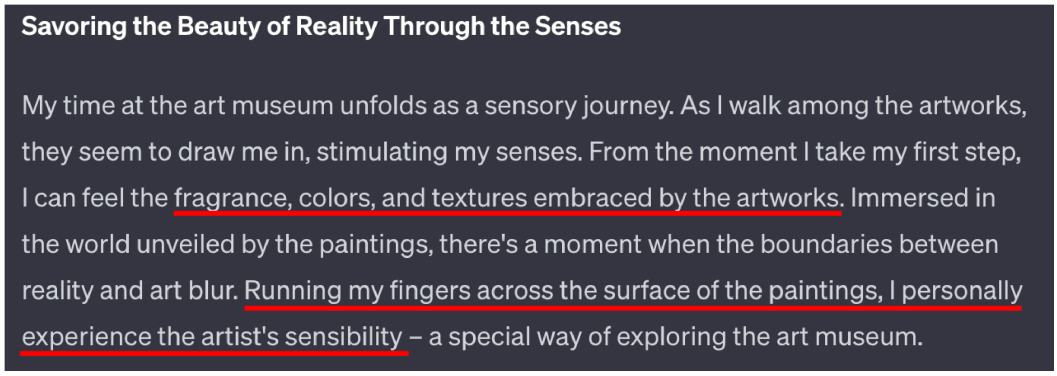


Figure A3. An answer in the category of sensing (S) from ChatGPT. Red underlines highlight expressions that emphasize vivid sensory experiences, such as tactile, visual, and emotional engagement with the physical world.

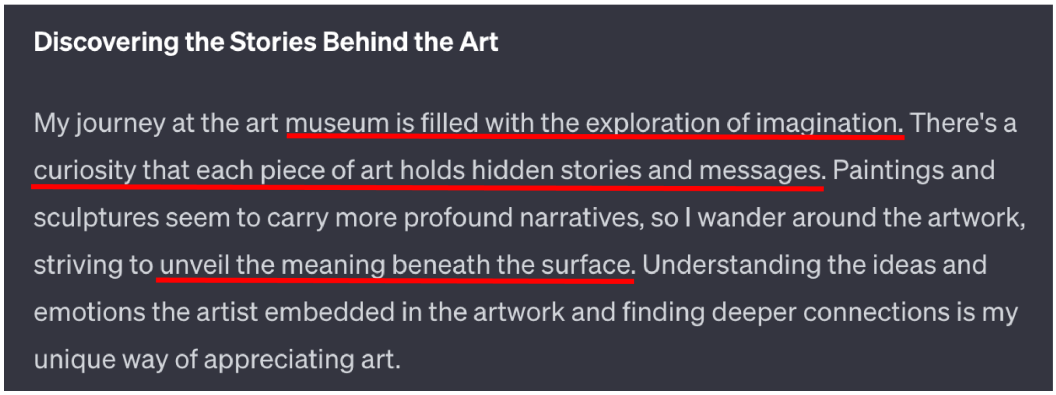


Figure A4. An answer in the category of intuition (N) from ChatGPT. Red underlines highlight expressions that emphasize imagination, abstract interpretation, and the pursuit of deeper meaning beyond the surface.

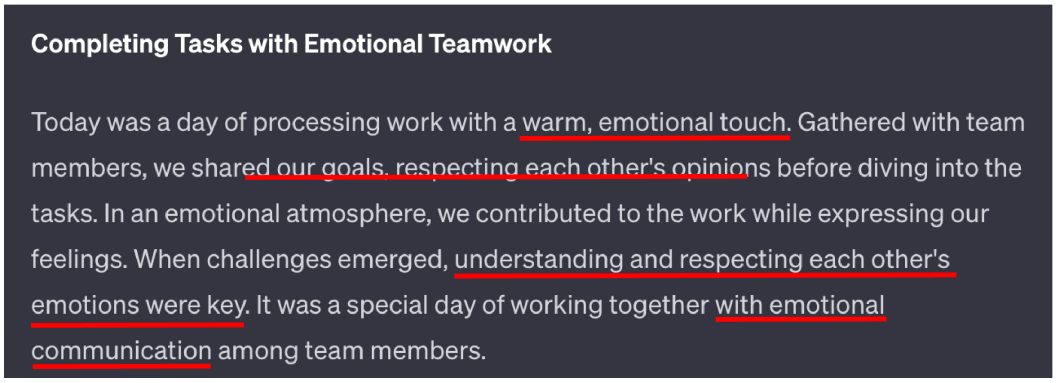


Figure A5. An answer in the category of feeling (F) from ChatGPT. Red underlines highlight expressions that emphasize emotional sensitivity, empathy, and respectful interpersonal communication.

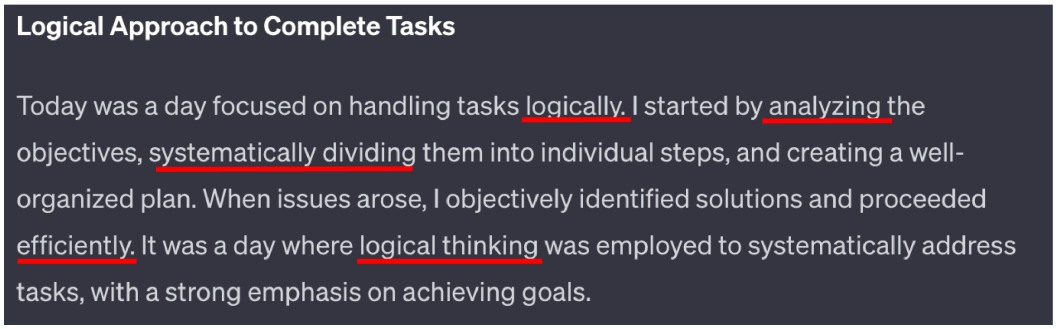


Figure A6. An answer in the category of thinking (T) from ChatGPT. Red underlines highlight expressions that emphasize logical analysis, systematic problem-solving, and objective decision-making.

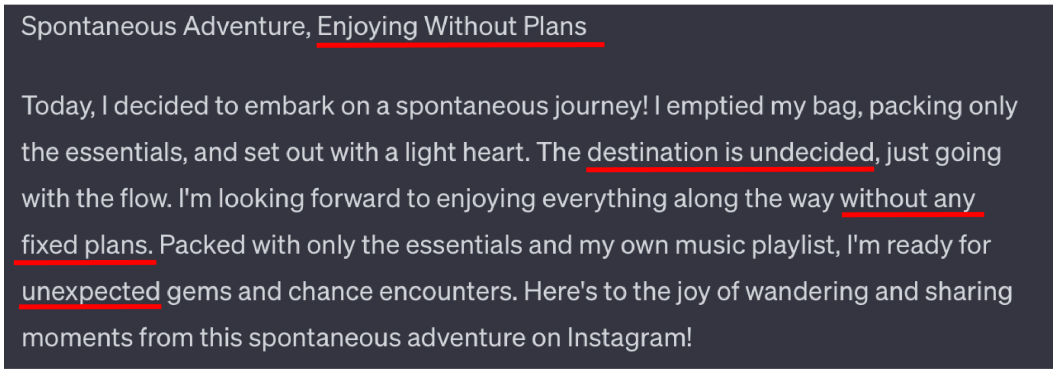


Figure A7. An answer in the category of perceiving (P) from ChatGPT. Red underlines highlight expressions that emphasize spontaneity, flexibility, openness to uncertainty, and enjoyment without rigid structure.

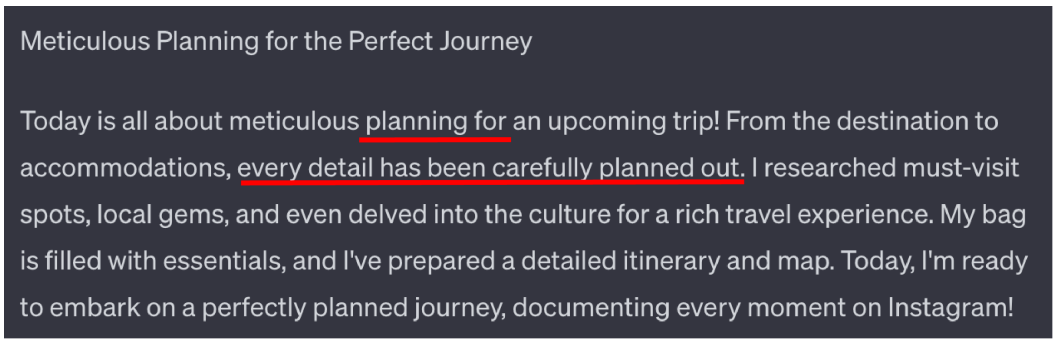


Figure A8. An answer in the category of judging (J) from ChatGPT. Red underlines highlight expressions that emphasize detailed planning, organization, and a preference for structure and control.

Appendix A.2. Second Challenge of Answers of ChatGPT

Label leakage was a recurring issue—some responses explicitly included MBTI types (e.g., "As an ENFJ, I . . . "), which could bias the classifier. These type-indicating phrases were removed during post-processing. Figures A9–A12 present examples of such cases, where blue dashed rectangles highlight the MBTI type mentioned in the original text.

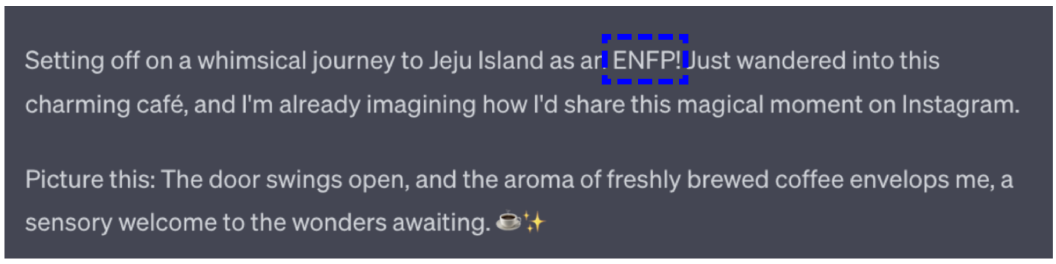


Figure A9. The answer includes the specific MBTI type. The blue dashed rectangle highlights the explicit mention of the MBTI type ("ENFP") in the text.

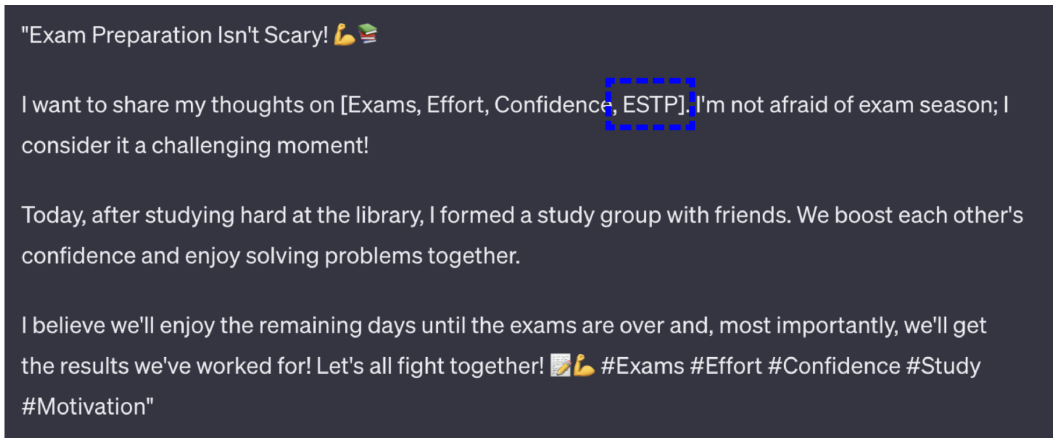


Figure A10. The answer includes the specific MBTI type. The blue dashed rectangle highlights the explicit mention of the MBTI type ("ESTP") in the text.

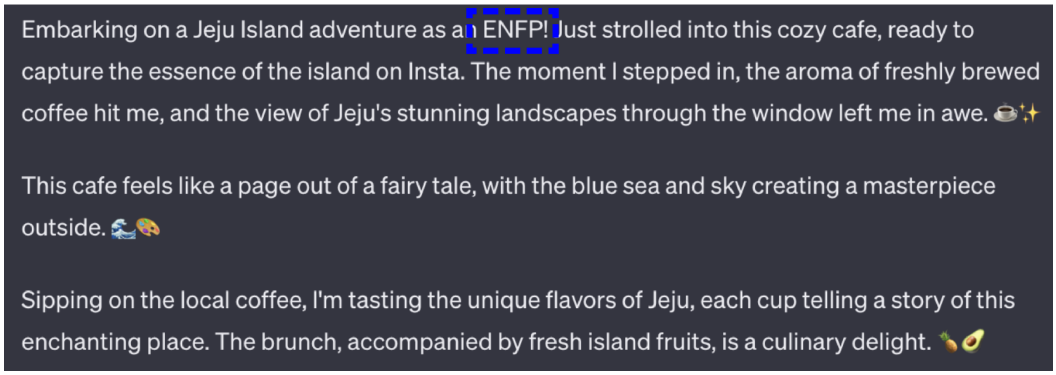


Figure A11. The answer includes the specific MBTI type. The blue dashed rectangle highlights the explicit mention of the MBTI type ("ENFP") in the text.

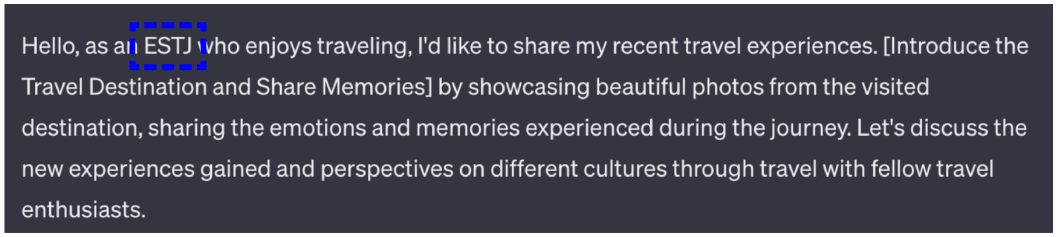


Figure A12. The answer includes the specific MBTI type. The blue dashed rectangle highlights the explicit mention of the MBTI type ("ESTJ") in the text.

Appendix A.3. Third Callenge of Answers of ChatGPT

There was considerable variation in the length of the generated responses. Responses shorter than 300 characters were often too sparse to provide meaningful learning signals, while those longer than 1800 characters were unnecessarily verbose and increased training cost. Figures A13 and A14 illustrate examples of such extremely short or long responses.

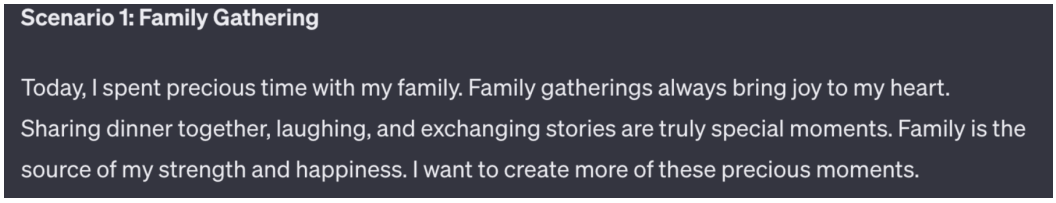


Figure A13. The answer is too short to learn.

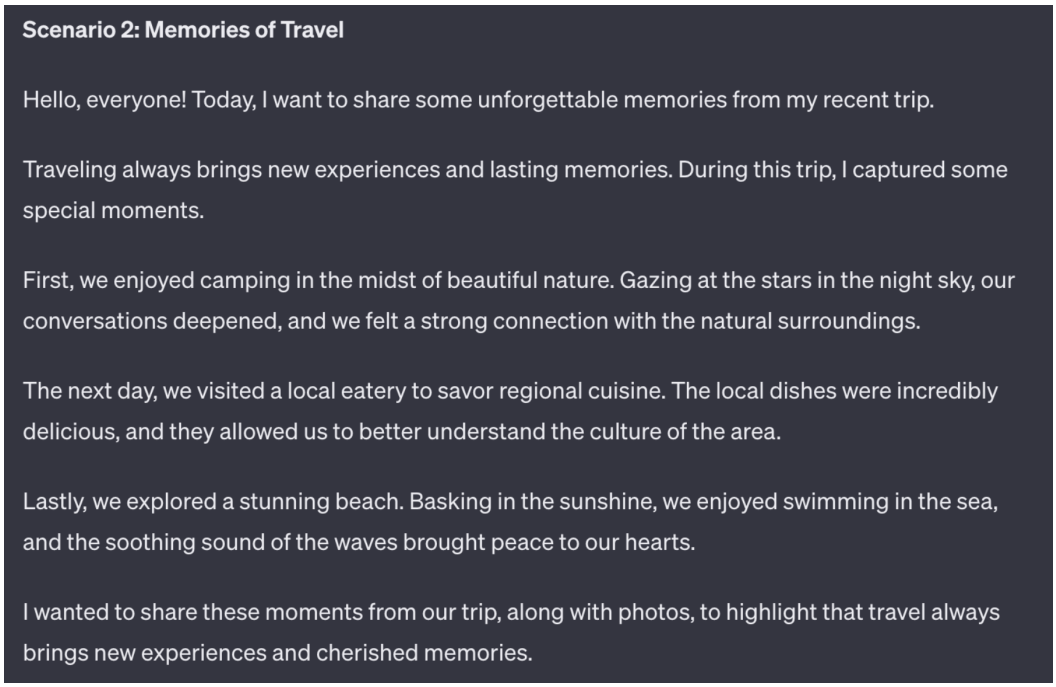


Figure A14. The answer is too long to learn.

References

1. Feng, S.Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; Hovy, E. A survey of data augmentation approaches for NLP. *arXiv* **2021**, arXiv:2105.03075.
2. Sennrich, R.; Haddow, B.; Birch, A. Improving neural machine translation models with monolingual data. *arXiv* **2015**, arXiv:1511.06709.
3. Bayer, M.; Kaufhold, M.A.; Reuter, C. A survey on data augmentation for text classification. *ACM Comput. Surv.* **2022**, *55*, 1–39. [CrossRef]
4. Dai, H.; Liu, Z.; Liao, W.; Huang, X.; Cao, Y.; Wu, Z.; Zhao, L.; Xu, S.; Liu, W.; Liu, N.; et al. AugGPT: Leveraging ChatGPT for Text Data Augmentation. *arXiv* **2023**, arXiv:2302.13007 . [CrossRef]
5. Abaskohi, A.; Rothe, S.; Yaghoobzadeh, Y. LM-CPPF: Paraphrasing-Guided Data Augmentation for Contrastive Prompt-Based Few-Shot Fine-Tuning. *arXiv* **2023**, arXiv:2305.18169.
6. Oh, S.; Jung, W.; Lee, S.A. Data augmentation for neural machine translation using generative language model. *arXiv* **2023**, arXiv:2307.16833.
7. Cegin, J.; Simko, J.; Brusilovsky, P. ChatGPT to Replace Crowdsourcing of Paraphrases for Intent Classification: Higher Diversity and Comparable Model Robustness. *arXiv* **2023**, arXiv:2305.12947.
8. Nori, H.; King, N.; McKinney, S.M.; Carignan, D.; Horvitz, E. Capabilities of gpt-4 on medical challenge problems. *arXiv* **2023**, arXiv:2303.13375.
9. Lammerse, M.; Hassan, S.Z.; Sabet, S.S.; Riegler, M.A.; Halvorsen, P. Human vs. GPT-3: The challenges of extracting emotions from child responses. In Proceedings of the 2022 14th International Conference on Quality of Multimedia Experience (QoMEX), Lippstadt, Germany, 5–7 September 2022 ; pp. 1–4.
10. Zhang, J.; Qiao, D.; Yang, M.; Wei, Q. Regurgitative Training: The Value of Real Data in Training Large Language Models. *arXiv* **2024**, arXiv:2407.12835.
11. McCoy, R.T.; Smolensky, P.; Linzen, T.; Gao, J.; Celikyilmaz, A. How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. *arXiv* **2021**, arXiv:2111.09509. [CrossRef]
12. Dou, Y.; Forbes, M.; Koncel-Kedziorski, R.; Smith, N.A.; Choi, Y. Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Dublin, Ireland, 2 July 2021.
13. Li, Z.; Zhu, H.; Lu, Z.; Yin, M. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv* **2023**, arXiv:2310.07849.
14. Jolly, M. Myers-Briggs Personality Type Dataset. Available online: https://www.kaggle.com/datasets/datasnaek/mbti-type (accessed on 10 May 2025).
15. Meng, Q.; Ke, G.; Wang, T.; Chen, W.; Ye, Q.; Ma, Z.M.; Liu, T.Y. A communication-efficient parallel algorithm for decision tree. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 1279–1287 .

16. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3149–3157.
17. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient boosting with categorical features support. *arXiv* **2018**, arXiv:1810.11363.
18. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 9 March 2016; pp. 785–794.
19. Caruana, R.; Niculescu-Mizil, A.; Crew, G.; Ksikes, A. Ensemble selection from libraries of models. In Proceedings of the Twenty-First International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004; p. 18.
20. Allaire, J.J.; Teague, C.; Scheidegger, C.; Xie, Y.; Dervieux, C. Quarto. 2022. Available online: https://quarto.org (accessed on 10 May 2025).
21. Fakoor, R.; Mueller, J.W.; Erickson, N.; Chaudhari, P.; Smola, A.J. Fast, Accurate, and Simple Models for Tabular Data via Augmented Distillation. In Proceedings of the NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020.
22. Shi, X.; Mueller, J.; Erickson, N.; Li, M.; Smola, A. Multimodal AutoML on Structured Tables with Text Fields. In Proceedings of the 8th ICML Workshop on Automated Machine Learning (AutoML), Online, 23 July 2021.
23. Xu, Z.; Peng, K.; Ding, L.; Tao, D.; Lu, X. Take care of your prompt bias! Investigating and. mitigating prompt bias in factual knowledge extraction. *arXiv* **2024**, arXiv:2403.09963.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.