# Detecting Primary Progressive Aphasia (PPA) from Text: A Benchmarking Study

**Ghofrane Merhbene**[1,2], **Fabian Lecron**[2], **Philippe Fortemps**[2],
**Bradford C. Dickerson**[3], **Mascha Kurpicz-Briki**[1], **Neguine Rezaii**[3]
[1]Bern University of Applied Sciences, Biel/Bienne, Switzerland
[2] University of Mons (UMONS), Mons, Belgium
[3] Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

ghofrane.merhbene@bfh.ch, fabian.lecron@umons.ac.be,
philippe.fortemps@umons.ac.be, brad.dickerson@mgh.harvard.edu,
mascha.kurpicz@bfh.ch, nrezaii@mgh.harvard.edu

## Abstract

Classifying subtypes of primary progressive aphasia (PPA) from connected speech presents significant diagnostic challenges due to overlapping linguistic markers. This study benchmarks the performance of traditional machine learning models with various feature extraction techniques, transformer-based models, and large language models (LLMs) for PPA classification. Our results indicate that while transformer-based models and LLMs exceed chance-level performance in terms of balanced accuracy, traditional classifiers combined with contextual embeddings remain highly competitive. Notably, MLP using MentalBert's embeddings achieves the highest accuracy. These findings underscore the potential of machine learning for enhancing the automatic classification of PPA subtypes.

## 1 Introduction

Primary progressive aphasia (PPA) is a neurodegenerative disorder characterized by progressive language deficits as the primary symptom. It is typically classified into three subtypes (Gorno-Tempini et al., 2011): (1) the logopenic variant (lvPPA), associated with word-finding difficulties and impaired sentence repetition, often linked to Alzheimer's pathology; (2) the semantic variant (svPPA), marked by deficits in word comprehension and object naming; and (3) the nonfluent variant (nfvPPA), characterized by effortful, halting, and telegraphic speech. The underlying pathology of svPPA and nfvPPA is often frontotemporal lobar degeneration (Rezaii et al., 2023). Diagnosing these subtypes traditionally requires extensive clinical assessment by expert neurologists, neuropsychologists, and speech-language pathologists, making the process resource-intensive and time-consuming. As a result, there is increasing interest in automated methods for efficient and accurate PPA classification. However, diagnosing PPA from textual data, such as transcripts of patient interviews, presents several challenges. The linguistic and syntactic markers that differentiate PPA subtypes are often subtle and overlapping, requiring robust feature extraction and classification techniques (Tippett, 2020). Furthermore, the limited availability of labeled clinical datasets and individual variability in language use exacerbate these challenges. Distinguishing svPPA from lvPPA is particularly difficult, as both subtypes involve word retrieval impairments. Despite these difficulties, accurate classification is crucial, given the distinct etiologies and treatment strategies associated with each PPA variant.

Recent advancements in natural language processing (NLP) have opened new avenues for automated diagnostic tools based on text. Prior research has demonstrated the potential of NLP in mental health assessment (Zhang et al., 2022), including applications in detecting bipolar disorder and schizophrenia (Aich et al., 2022). However, research on applying NLP to neurodegenerative diseases, particularly PPA, remains limited. Notably, there is a lack of systematic benchmarking studies that compare multiple computational approaches for PPA classification. To address this gap, we conduct a comprehensive benchmarking study, systematically evaluating a diverse range of models, from traditional machine learning (ML) methods with various feature extraction techniques to transformer-based models and large language models (LLMs). By providing a comparative analysis of these approaches, our study offers new insights into the effectiveness



*Findings of the Association for Computational Linguistics: EACL 2026*, pages 355–374
March 24-29, 2026 

of different computational techniques for the automated classification of PPA subtypes.

## 2 Related Work

Research on PPA has primarily focused on understanding its clinical subtypes and linguistic manifestations (Henry et al., 2016). Studies in clinical neurology and neuropsychology have detailed the unique language impairments associated with lvPPA, svPPA, and nfvPPA, highlighting the importance of linguistic and syntactic analysis in diagnosis (Wauters et al., 2023). However, leveraging computational methods for the diagnosis of PPA remains an emerging area.

In the field of natural language processing (NLP), traditional ML models have been widely applied to clinical text classification tasks, including disease detection and subtype identification. In their study, Fraser et al. (2014) explored the use of computational linguistics for identifying different variants of PPA. They compared various feature sets, including acoustic, lexical, and syntactic features, and demonstrated that combining multiple modalities significantly improved classification performance. Their findings highlighted the importance of leveraging diverse linguistic markers to distinguish PPA subtypes, particularly the nonfluent variant (nfvPPA), which often exhibits clear syntactic deficits. Similarly, Themistocleous et al. (2021) achieved a classification accuracy of 80% by combining acoustic and linguistic features and using them as input for a deep neural network model. Building on this foundation, Rezaii et al. (2022) investigated the relationship between lexical and syntactic complexity during language production in individuals with PPA and healthy controls. Their study identified a syntax-lexicon trade-off where individuals with syntactic deficits, such as those with nfvPPA, used semantically richer words, while those with lexicosemantic deficits (e.g., svPPA or lvPPA) produced syntactically complex sentences. Their approach achieved a classification accuracy of up to 92% when distinguishing nfvPPA in a one-vs-all setup. In more recent work, Rezaii et al. (2024) explored the use of LLMs to classify PPA subtypes based on connected speech. Their approach incorporated verb frequency and other linguistic features to align text-based speech patterns with brain scan findings, achieving 88.5% agreement on PPA clusters with LLMs. A supervised classifier using features identified by the LLM further improved accuracy to 97.9%. This study highlights the potential of LLMs in identifying linguistic markers of PPA subtypes and represents a significant advance in the application of NLP to clinical tasks. Cong et al. (2024b) also investigated the use of LLMs for detecting the presence, subtypes, and severity of aphasia in both English and Mandarin Chinese speakers. Their findings revealed that applying LLMs without fine-tuning resulted in accuracy levels close to chance for aphasia subtyping.

Language impairments, such as PPA, are often among the earliest signs of broader cognitive decline, including dementia (Harvard Health Publishing, 2022). Santander-Cruz et al. (2022) employed a combination of syntactic and semantic analyses to detect dementia in transcribed data from the Pitt Corpus database provided by DementiaBank[1]. They extracted features such as spelling mistakes, grammar errors, and cosine similarity and evaluated their effectiveness using ML models, including SVMs and neural networks. Notably, syntactic features alone achieved an F1-score of 77% with SVMs. While their approach demonstrated the effectiveness of syntactic features, it remained limited in scope, focusing on a predefined feature set and a small selection of models. In contrast, our study systematically evaluates a wider range of methodologies, from traditional ML models with different feature extraction techniques to transformer-based models and LLMs, to comprehensively assess the potential of NLP techniques for PPA classification.

## 3 Dataset

### 3.1 Overview

The data used in this study were shared with us by Rezaii et al. (2022) and made available by the authors on the Open Science Framework. The dataset consists of clinical transcripts from interviews with individuals diagnosed with one of the PPA subtypes, as well as control participants without a PPA diagnosis. A key limitation of text-based analyses is that public sharing of voice data remains restricted due to concerns about participant identification. However, an advantage of this work is that patients can still be classified based on their written texts (e.g., Josephy-Hernandez et al. (2023)). Further details about the data are provided in Appendix A.

[1] https://dementia.talkbank.org/

Two versions of the dataset were used in this study: the *original version*, which includes each participant's full interview transcript, and the *expanded version*, where each transcript was split into individual sentences, with each sentence inheriting the label of the original transcript. The label distribution for both versions of the dataset is provided in Tables 3 and 4 in Appendix A.

Statistics for both versions of the dataset, including the mean, median, and standard deviation of text lengths in words, are presented in Table 5 in Appendix A.

### 3.2 Data Preprocessing

Preprocessing the data is a crucial step before applying ML models, as it ensures the integrity of the linguistic and syntactic features. This section details the preprocessing steps undertaken to prepare the dataset.

The first step included converting all text to lowercase to standardize case sensitivity. Special characters were removed, retaining only intentionally included alphanumeric characters and punctuation marks, as these features are significant in the diagnosis of PPA. For instance, punctuation patterns can signify pauses, sentence boundaries, or telegraphic speech, which are critical markers for distinguishing between PPA subtypes. nfvPPA, in particular, is marked by halting speech and frequent pauses. Following this, the text was tokenized into individual words for further analysis.

It is important to mention that the preprocessing steps were applied for the experiments with the traditional ML models described in Section 4.2.

## 4 Methodology

The code used in this study is made publicly available at GitHub link.

### 4.1 Evaluation Reference Points

To evaluate the performance of the models in this multi-class classification task, we define a reference metric to provide a point of comparison for balanced accuracy:

**Stratified (Weighted) Random Reference:** This reference metric accounts for class imbalance by weighting each class proportionally to its frequency in the dataset. Since this metric incorporates dataset imbalance, it provides a more realistic reference than uniform random guessing.

$$\sum_{i=1}^{N} P(\text{Class}_i)^2 = \sum_{i=1}^{N} \left( \frac{\text{ClassCount}_i}{\text{TotalSamples}} \right)^2 \quad (1)$$

### 4.2 Traditional Machine Learning Models

The initial experiments in this benchmarking study involve applying various feature extraction techniques in combination with a predefined set of traditional ML models. The following subsection provides an overview of the feature extraction techniques used.

#### 4.2.1 Feature Extraction techniques

Several feature extraction strategies were evaluated in this study, spanning from traditional statistical methods to more advanced embedding-based and syntactic techniques. TF-IDF (Salton and Buckley, 1988) and Bag-of-Words (BoW) (Harris, 1954) focused on capturing word frequency and document-level term relevance. To incorporate semantic information, we employed embedding-based models such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017); the latter also accounts for subword structures. For contextual representation, we extracted embeddings from transformer-based models including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), MentalBERT (Ji et al., 2022), and ClinicalBERT (Alsentzer et al., 2019). Additional features were derived using N-grams (from bigrams to 4-grams) (Brown et al., 1992) to capture local context, LSA (Deerwester et al., 1990) and LDA (Blei et al., 2003) for latent topic modeling, and dependency parsing (Kiperwasser and Goldberg, 2016) to model syntactic relationships.

#### 4.2.2 Machine Learning Models

Traditional ML models have played a key role in advancing AI and continue to offer advantages such as interpretability, computational efficiency, and adaptation to smaller datasets (Murphy, 2012). Despite the growing dominance of LLMs, the performance of traditional models should not be overlooked, particularly in tasks where linguistic and syntactic features play a central role.

To ensure a robust benchmarking process, we incorporate five widely-used traditional ML models: Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR), Multilayer Perceptron (MLP), and XGBoost. These models were evaluated in combination with the feature extraction techniques detailed in the previous section.

The *expanded* version of the dataset was used for this experiment. The decision to split the *original* dataset at the sentence level was motivated by the goal of aligning with a prior study that used the same dataset to ensure comparability, as well as to increase the number of training examples.
Considering the limited sample size (79 PPA patients, 53 controls), default parameter settings without hyperparameter fine-tuning were used for all models. This conservative approach minimizes overfitting risk inherent in tuning hyperparameters on small datasets and ensures fair comparison across models by avoiding model-specific advantages from optimization. While this may not reflect the maximum achievable performance for any individual model, it provides a reproducible baseline.
To prevent data leakage, feature extraction was integrated within scikit-learn pipelines, ensuring that feature computation was performed solely on the training data during each fold and never on the test data. Additionally, `GroupKFold` cross-validation was used to ensure that all data from a single participant appeared exclusively in either the training folds or the test fold, thereby preventing data leakage across splits. This prevented the model from learning to recognize individual participants instead of the targeted PPA subtype. In total, 65 experiments were conducted (5 classifiers × 13 feature extraction techniques).

### 4.3 Transformer-based Models

In addition to traditional ML models, this study evaluates the performance of transformer-based models, which have revolutionized natural language processing by taking advantage of attention mechanisms and contextual embeddings. These models are particularly well-suited for tasks involving subtle syntactic variations and capturing long-term dependencies, making them strong candidates for the classification task at hand. While some transformer models were previously used to generate embeddings for feature-based approaches (as detailed above), here, they are directly employed as classifiers to assess their full predictive capabilities.

The transformer-based models included in this benchmarking study are as follows: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), MentalBERT (Ji et al., 2022), and ClinicalBERT (Alsentzer et al., 2019). A detailed description of each model is provided in Appendix B.

These models are evaluated using the same cross-validation protocols applied to traditional ML models, ensuring fair comparison. Each training involved re-initializing the model and optimizer, followed by full fine-tuning for 10 epochs on the training split.
The dataset exhibits a moderate class imbalance (see Table 4). Since this work presents a benchmarking study where both traditional ML and transformer-based classifiers are evaluated under the same cross-validation settings without additional resampling or weighting techniques, no explicit method for addressing class imbalance (e.g., class weights or oversampling) was applied. This consistent protocol allows for fair comparisons across model types. However, we acknowledge that class imbalance may still impact the performance of some classifiers, especially on underrepresented subtypes.

## 5 Large Language Models (LLMs)

LLMs represent a significant breakthrough in artificial intelligence, demonstrating exceptional capabilities across a wide range of NLP tasks. These models, like OpenAI's GPT series and Google's Gemini, are built upon transformer-based architectures and are known by their immense size, comprising billions or even trillions of parameters. Their extensive training, combined with their parameterization, allows them to achieve high performance in a wide range of NLP tasks, including text generation.

In this study, we employ a prompt-based approach to leverage LLMs for the classification of PPA subtypes. Rather than fine-tuning these models, we evaluate their zero-shot performance by designing a structured prompt tailored to our classification task. The following LLMs were used in this study: LLAMA (Touvron et al., 2023), Mistral (Jiang et al., 2023), GPT-3.5-turbo (Brown et al., 2020), and GPT-4o-mini (OpenAI, 2023). Detailed descriptions of each model are provided in Appendix B.
The *original* version of the data was used, and the prompt was carefully designed in collaboration with a clinical expert in the field (see Appendix C). The temperatures used for each model are presented in Table 1. For Mistral and LLAMA, we used a relatively low temperature (0.2) to ensure more deterministic outputs[2], as these models may

[2] https://huggingface.co/docs/transformers/mai

exhibit greater output variability at higher temperatures. In contrast, GPT-3.5 and GPT-4o-mini were assigned a moderately higher temperature (0.7) to encourage more diverse responses while maintaining overall coherence. This choice was informed by prior observations that hallucination rates tend to be higher in open-source models such as LLAMA and Mistral, and that lower temperatures help mitigate this issue (Yang et al., 2025).

| Model | Temperature |
|---|---|
| Mistral | 0.2 |
| LLAMA | 0.2 |
| GPT-3.5 | 0.7 |
| GPT-4o-mini | 0.7 |

Table 1: Temperature used for each model.

# 6 Results

To ensure a comprehensive evaluation, we rely on widely recognized classification metrics, including balanced accuracy, weighted F1-score, weighted precision, weighted recall, Area Under the Curve (AUC), as well as a confusion matrix for LLM based experiments. All experiments are evaluated using 5-fold cross-validation to ensure robustness and minimize overfitting. The results are presented as bar charts, with balanced accuracy's reference performance indicated by vertical lines to provide a clear point of comparison. Additionally, local feature importance analyses were conducted using LIME (Ribeiro et al., 2016) for the top-performing models in both the traditional ML and transformer-based experiments, providing insight into which input features most influenced individual predictions.

## 6.1 Traditional Machine Learning Models

Figure 1 presents the performance of the top-performing traditional ML models (in terms of F1-score), namely MLP. Each colored bar represents the ML model paired with a different feature extraction technique. The results for the other models, including LR, SVM, NB, and XGBoost, are provided in Appendix E for completeness.

In terms of balanced accuracy, features derived from MentalBERT, followed by those from BERT, consistently yielded the best results across nearly all models. LR showed comparable performance when using MentalBERT, BERT, and Bag-of-Words features. MentalBERT also outperformed other models across additional metrics, including weighted precision, weighted recall, weighted F1-score, and AUC, with BERT and RoBERTa following closely. Notably, MentalBERT achieved over 60% on weighted precision, recall, and F1-score for the MLP classifier, and reached or approached 80% AUC with MLP, SVM, LR, and XGBoost.

## 6.2 Transformer-based Models

Figure 2 illustrates the performance of the various transformer-based classifiers. All models significantly outperform the reference metric in terms of balanced accuracy, with RoBERTa and BERT demonstrating comparable top-tier performance, closely followed by MentalBERT. Regarding the F1-score, RoBERTa and BERT achieve the highest results of 57%. Similar trends are observed for weighted precision and weighted recall, where RoBERTa and BERT achieve scores a little under 60%. In terms of AUC, RoBERTa, BERT and MentalBERT all demonstrate strong performance, achieving results at or near 80%.

## 6.3 Large Language Models (LLMs)

Figure 3 presents a bar chart illustrating the performance of LLAMA, which achieved the highest weighted precision, weighted recall, and F1-score among all LLMs. In terms of balanced accuracy, it was outperformed only by GPT-4o-mini. For completeness, the results of Mistral, GPT-3.5-turbo and GPT-4o-mini are provided in Appendix F. All models, except for Mistral, outperformed our reference metric.

Figures 11, 12, 13, and 14 in Appendix F present the confusion matrices for all four models. For LLAMA, we observe a strong performance in correctly predicting both the control group and lvPPA, but the model struggled with predicting any svPPA samples. Mistral's performance, shown in Figure 12, was the weakest, as it assigned multiple times the label, *unknown*, when it failed to classify a sample correctly. Both GPT models performed well in identifying the control group, with GPT-3.5-Turbo showing a slight edge over the other model. However, both models faced significant difficulty with lvPPA. GPT-3.5 also showed limited success with svPPA, whereas GPT-4o-mini performed better on both svPPA and nfvPPA.

n_classes/text_generation#parameters

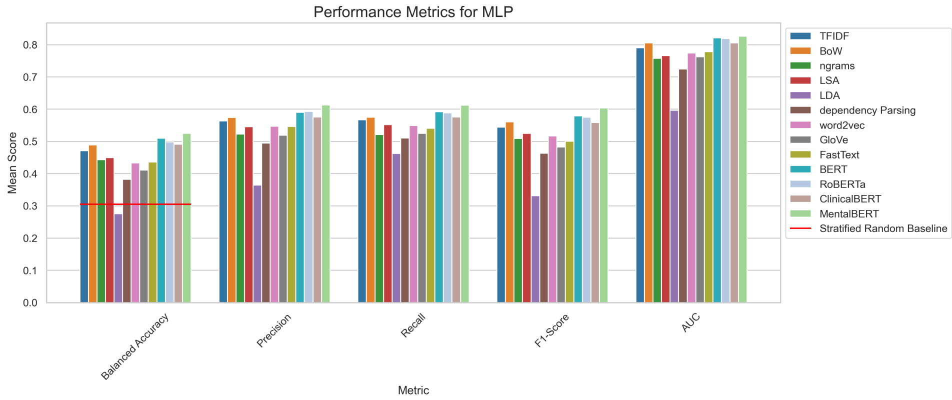


Figure 1: MLP performance


Figure 2: Transformer-based models performance

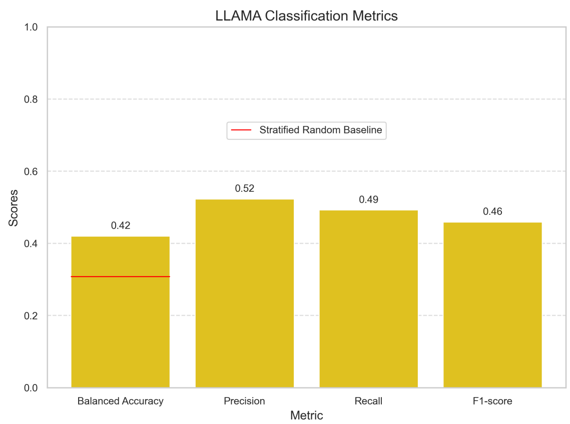


Figure 3: LLAMA Performance

## 7 Discussion

The results of this study provide valuable insights into the potential of various models for detecting primary progressive aphasia (PPA) subtypes. Benchmarking traditional ML approaches, transformer-based models, and LLMs holds significant importance in advancing clinical diagnostics. These efforts not only reveal key trends and performance disparities but also underscore the broader potential of these models to improve the detection and classification of complex clinical conditions, such as PPA.

The results from traditional ML models reveal the critical role of feature extraction in determining performance. In particular, embeddings derived from transformer-based models such as MentalBERT, RoBERTa, and BERT consistently outperformed classical feature engineering methods across most classifiers. This was especially evident in MLP, where the use of MentalBERT's features resulted in reaching or exceeding 60% weighted precision, weighted recall, and F1-scores, as well as AUC values exceeding 80%. These findings highlight the potential of combining robust feature extraction methods with simpler classifiers to achieve competitive results, especially in resource-constrained environments. In addition, in use cases where context is important relying on contextual embeddings like those generated by transformer-based models is generally expected to yield better results. The LR model paired with BoW features still demonstrated competitive results, closely trailing behind transformer-based embeddings. This further suggests that simpler techniques may still be viable in certain scenarios, particularly when interpretability is prioritized (Itani et al., 2019). When dealing with sensitive medical conditions such as PPA, interpretation is paramount, as clinicians and researchers need to understand the rationale behind model predictions. The ability to explain why a model classified a patient's condition can thus foster trust.

In line with recent work by Rezaii et al. (2022), our findings further emphasize the inherent difficulty of the multi-class classification task for PPA subtypes. Relying on a syntax-lexicon approach, the authors achieved a high accuracy (92%) in a binary classification task but reported a significant drop to 66% accuracy in multi-class classification. This stark contrast underscores the challenges faced by the overlapping symptoms and complexity of different PPA subtypes. Similar to their findings, our results confirm that advanced ML techniques, while promising, still face limitations when addressing multi-class classification in this domain.

Furthermore, transformer-based models such as RoBERTa and BERT achieved balanced accuracy and F1-scores 57%, which highlights the intrinsic challenges of capturing the subtle linguistic and syntactic variations inherent in PPA subtypes in a multi-class classification setting. These results align with the broader challenges outlined by Gorno-Tempini et al. (2011), who discussed the diagnostic complexity of PPA due to the heterogeneity and overlapping symptoms among its subtypes. While transformer models demonstrated promising results, they were outperformed by traditional ML models combined with transformer-based embeddings. This suggests that although transformers hold potential for capturing complex linguistic patterns, further refinement and task-specific adaptation are necessary to fully leverage their capabilities. This finding was also emphasized by Cong et al. (2024a), where the authors reaffirmed the potential of transformer-based models in healthcare, particularly in identifying complex patterns essential for the early detection and classification of neurodegenerative diseases. In addition, an important insight is that general-domain models appear to outperform domain-specific ones. Specifically, RoBERTa and BERT consistently produced stronger results than ClinicalBERT and MentalBERT, although MentalBERT's performed comparably on most metrics. One possible explanation is that larger, more diverse pretraining corpora may help general-domain models capture a wider range of linguistic cues. However, even if domain-specific models are adjusted to specialised vocabulary, they could overlook some contextual cues or universal language patterns that are useful in broader tasks. In fact, general-domain BERT can occasionally stay competitive or even outperform specialised models, according to Alsentzer et al. (2019), indicating that in some situations, greater coverage may outweigh niche specialization in certain scenarios.

Although BERT and RoBERTa achieve the best scores among the end-to-end transformer models, they are still outperformed by a lighter pipeline in which a frozen MentalBERT encoder feeds an MLP classifier. This gap can be explained by two factors. First, MentalBERT is pre-trained on clinical and mental-health text, so its embeddings inherently capture stylistic cues like telegraphic phrases, disfluencies, domain vocabulary (that are highly relevant to PPA), whereas generic BERT/RoBERTa must learn these patterns from the small fine-tuning set. Second, full fine-tuning updates hundreds of millions of parameters and is prone to overfitting when data are limited and moderately imbalanced (Devlin et al., 2019).

While LLMs outperformed our reference metric in terms of balanced accuracy (with the exception of Mistral), their results were inconsistent across the subtypes (see confusion matrices in Appendix F). LLAMA achieved the highest weighted precision, weighted recall, and F1-score, yet it struggled particularly with svPPA classification. One likely explanation is that we used these models without fine-tuning, relying solely on prompting. Unlike smaller models explicitly optimized for classification through feature-based learning, LLMs generate responses based on broad language modeling objectives, which may not align well with structured clinical classification. These results highlight the limitations of zero-shot LLM classification, where performance may be constrained without fine-tuning or domain adaptation. Table 2 highlights the best-performing models across our experiments. While most models demonstrated comparable performance, LLAMA stood out negatively; despite outperforming other LLMs, it failed to match the top models in other categories. MLP paired with MentalBERT's embeddings emerged as the strongest model, achieving the highest scores in balanced accuracy, weighted F1-score, weighted precision, and weighted recall, though by a narrow margin.

| Model | Bal. Acc. | F1 | P | R |
|---|---|---|---|---|
| LLAMA | 0.42 | 0.46 | 0.52 | 0.49 |
| BERT | 0.51 | 0.57 | 0.60 | 0.58 |
| RoBERTa | 0.51 | 0.57 | 0.60 | 0.57 |
| MLP & MentalBERT | **0.52** | **0.60** | **0.61** | **0.61** |

Table 2: Performance metrics for the best models (Bal. Acc. = Balanced Accuracy, P = Precision, R = Recall). Highest value(s) in each column are in bold.

Additionally, we use LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016) to analyze local feature importance for individual predictions from our best-performing model (MLP with MentalBERT embeddings), as well as BERT and RoBERTa. We present one example per subtype in Appendix G. For svPPA (see Figures 15, 16, and 17), non-specific words like *people* consistently received high importance across all three models, aligning with known svPPA speech patterns (Gorno-Tempini et al., 2011). Similarly, frequent verbs such as *sitting*, *getting*, and *eating* were among the most influential tokens, which is also characteristic of svPPA language use (Lukic et al., 2022). The word *two* was weighted negatively, indicating Not svPPA, which aligns with the observation that svPPA patients tend to use vague and general language rather than specific quantifiers (Faust et al., 2012). In the case of lvPPA (see Figures 18, 19, and 20), patients often use interjections and fillers to mask disfluencies such as *uhh*, which received notable importance, particularly in the MLP + MentalBERT model. Indefinite determiners like *a* were assigned the highest importance by both MLP + MentalBERT and BERT, which aligns with the findings of (Robertson et al., 2024) and reflects the lexical retrieval difficulties typical of lvPPA. In contrast, RoBERTa did not highlight these tokens as strongly, which may be due to differences in pretraining data or tokenization. Notably, the filler *uhh* was deliberately transcribed in a specific way that may not align with RoBERTa's subword vocabulary, limiting its interpretability. For nfvPPA (see Figures 21, 22, and 23), the use of content nouns like *girl* was consistently highlighted across the three models, aligning with known speech patterns of nfvPPA patients. Interestingly, the word *sanding* -which is not a real word in this context and was invented by the patient— received the highest importance in BERT. This may reflect BERT's sensitivity to surface morphology, particularly -ing endings, which are frequently used by nfvPPA patients (Wilson et al., 2010). In contrast, *sanding* was negatively weighted by MLP and RoBERTa, while a concrete noun like *castle* was ignored only by BERT. These inconsistencies highlight the models' differing sensitivities and suggest that integrating their complementary perspectives may lead to more robust and clinically meaningful interpretations in future work.

## 8 Conclusion

Our findings show the promise of using ML in the classification of PPA subtypes. The results demonstrate that although transformer-based methods sometimes yield comparable metrics, they do not decisively outperform classical feature based techniques such as MLP paired with MentalBERT's embeddings. This highlights the inherent complexity of the classification task, shaped by the overlapping symptoms across PPA subtypes. Given the limitations observed in prompt-based LLM experiments, future work should explore task-specific fine-tuning to better align these models with the linguistic characteristics of PPA. Further error analysis may also provide insights into systematic misclassifications, guiding refinements in model training.

## 9 Limitations

The task of classifying primary progressive aphasia (PPA) subtypes presents a significant challenge due to the overlapping symptoms and linguistic impairments between subtypes. Additionally, our dataset, while useful for benchmarking remains relatively small and lacks demographic metadata, preventing an analysis of potential biases across different population groups. Computational constraints also limited our ability to explore hyperparameter tuning for all our experiments, which may have impacted model performance. This is particularly relevant for traditional classifiers and transformer-based models, where optimal settings could have led to improved results. Similarly, our exclusive reliance on natural language prompts for LLMs (although designed with expert input) may have limited their performance, as we lacked fine-tuning or deeper insights into their decision-making processes. The small dataset size also limits our ability to fully leverage the potential of LLMs, which typically benefit from larger-scale training or adaptation data. Without explicit control over how LLMs generate classifications, their outputs can be difficult to interpret and optimize for this task. Future work should explore fine-tuning approaches and systematic hyperparameter optimization to better align model performance with the complexities of PPA classification. Additionally, it is generally recommended to repeat LLM-based experiments and report average performance along with standard deviations, especially given the models' non-deterministic nature and the small size of our dataset. However, this was not feasible in our case due to limited computational resources.

Additionally, our classification approach relies solely on textual data. While this enables certain forms of linguistic analysis, it overlooks crucial acoustic features that are particularly relevant in the context of Primary Progressive Aphasia (PPA), where speech characteristics such as pronunciation, pause duration, and stuttering play a significant diagnostic role. Unfortunately, due to data privacy constraints, access to audio recordings or transcriptions was not possible in our study.

## 10 Ethical Considerations

The dataset used in this research was anonymized and sourced from a prior work. This ensures that the privacy and data protection of the original participants are upheld. However, due to the anonymization process, we have limited information about participants' demographic backgrounds. As a result, we cannot assess potential biases or limitations of our classifiers across different societal groups. To ensure broader applicability and fairness, it is essential to validate our findings on a larger and more diverse dataset before considering real-world deployment.

Additionally, while this work does not directly create an automated diagnostic tool, its findings could contribute to the development of such technologies in the future. We emphasize that the goal is to assist clinicians rather than replace them, and we acknowledge the potential risk of misuse if such tools were to be used as substitutes for expert judgment.

## References


Ankit Aich, Avery Quynh, Varsha Badal, Amy Pinkham, Philip Harvey, Colin Depp, and Natalie Parde. 2022. Towards intelligent clinically-informed language analyses of people with bipolar disorder and schizophrenia. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2871–2887, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1992. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(4):467–472.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Heather M. Clark, Rene L. Utianski, Joseph R. Duffy, Edythe A. Strand, Herholtz Botha, Keith A. Josephs, and Jennifer L. Whitwell. 2020. Western aphasia battery-revised profiles in primary progressive aphasia and primary progressive apraxia of speech.

*American Journal of Speech-Language Pathology*, 29(1S):498–510.

Shan Cong, Hang Wang, Yang Zhou, Zheng Wang, Xiaohui Yao, and Chunsheng Yang. 2024a. Comprehensive review of transformer-based models in neuroscience, neurology, and psychiatry. *Brain-X*, 2(2):e57.

Yan Cong, Jiyeon Lee, and Arianna LaCroix. 2024b. Leveraging pre-trained large language models for aphasia detection in English and Chinese speakers. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 238–245, Mexico City, Mexico. Association for Computational Linguistics.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Miriam Faust, Elisheva Ben-Artzi, and Nili Vardi. 2012. Semantic processing in native and second language: Evidence from hemispheric differences in fine and coarse semantic coding. *Brain and Language*, 123(3):228–233.

Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. 2014. Comparison of different feature sets for identification of variants in primary progressive aphasia. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 17–26, Baltimore, Maryland, USA. Association for Computational Linguistics.

Maria Luisa Gorno-Tempini, Argye E Hillis, Sandra Weintraub, Andrew Kertesz, Mario Mendez, Stefano F Cappa, Jennifer M Ogar, Jonathan D Rohrer, Sandra Black, Bradley F Boeve, et al. 2011. Classification of primary progressive aphasia and its variants. *Neurology*, 76(11):1006–1014.

Zellig S. Harris. 1954. Distributional structure. *WORD*, 10(2-3):146–162.

Harvard Health Publishing. 2022. Progressive aphasia involves many losses—here's what you need to know. Accessed: 2025-02-11.

Maya L. Henry, Stephen M. Wilson, Mary C. Babiak, Maria L. Mandelli, Pélagie M. Beeson, Zoe A. Miller, and Maria Luisa Gorno-Tempini. 2016. Phonological processing in primary progressive aphasia. *Journal of Cognitive Neuroscience*, 28(2):210–222.

Sarah Itani, Fabian Lecron, and Philippe Fortemps. 2019. Specifics of medical data mining for diagnosis aid: A survey. *Expert Systems with Applications*, 118:300–314.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Sylvia Josephy-Hernandez, Negar Rezaii, Amanda Jones, Elise Loyer, David Hochberg, Meaghan Quimby, Bruce Wong, and Bradford C. Dickerson. 2023. Automated analysis of written language in the three variants of primary progressive aphasia. *Brain Communications*, 5(4):fcad202.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Sladjana Lukic, Abigail E. Licata, Elizabeth Weis, Rian Bogley, Buddhika Ratnasiri, Ariane E. Welch, Leighton B. N. Hinkley, Z. Miller, Adolfo M. Garcia, John F. Houde, Srikantan S. Nagarajan, Maria Luisa Gorno-Tempini, and Valentina Borghesani. 2022. Auditory verb generation performance patterns dissociate variants of primary progressive aphasia. *Frontiers in Psychology*, Volume 13 - 2022.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Kevin P. Murphy. 2012. Machine learning - a probabilistic perspective. In *Adaptive computation and machine learning series*.

OpenAI. 2023. Gpt-4o-mini: Advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`. Accessed: 2025-02-06.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Neguine Rezaii, Daisy Hochberg, Megan Quimby, Bonnie Wong, Michael Brickhouse, Alexandra Touroutoglou, Bradford C. Dickerson, and Phillip Wolff. 2024. Artificial intelligence classifies primary progressive aphasia from connected speech. *Brain*, 147(9):3070–3082.

Neguine Rezaii, Kyle Mahowald, Rachel Ryskin, Bradford Dickerson, and Edward Gibson. 2022. A syntax-lexicon trade-off in language production. *Proceedings of the National Academy of Sciences*, 119(25):e2120203119.

Neguine Rezaii, Boyu Ren, Megan Quimby, Daisy Hochberg, and Bradford C. Dickerson. 2023. Less is more in language production: an information-theoretic analysis of agrammatism in primary progressive aphasia. *Brain Communications*, 5.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

C. Robertson, N. Rezaii, D. Hochberg, M. Quimby, P. Wolff, and B. C. Dickerson. 2024. Using explainable artificial intelligence to identify linguistic biomarkers of amyloid pathology in primary progressive aphasia. *medRxiv*. Preprint.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Yamanki Santander-Cruz, Sebastián Salazar-Colores, Wilfrido Jacobo Paredes-García, Humberto Guendulain-Arenas, and Saúl Tovar-Arriaga. 2022. Semantic feature extraction using sbert for dementia detection. *Brain Sciences*, 12(2):270.

Charalambos Themistocleous, Bronte Ficek, Kimberly Webster, Dirk-Bart den Ouden, Argye E. Hillis, and Kyrana Tsapkini. 2021. Automatic subtyping of individuals with primary progressive aphasia. *Journal of Alzheimer's Disease*, 79(3):1185–1194.

Donna C. Tippett. 2020. Classification of primary progressive aphasia: challenges and complexities. *F1000Research*, 9:F1000 Faculty Rev–64.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Awanish Batra, Daniel Haziza, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2307.09288*.

Lisa Wauters, Karen Croot, Heather Dial, Joseph R. Duffy, Stephanie M Grasso, Esther Kim, Kristin M. Schaffer, Kirrie J. Ballard, Heather M. Clark, Leeah Kohley, Laura L. Murray, Emily Rogalskı, Mathieu Figeys, Lisa H. Milman, and Maya L. Henry. 2023. Behavioral treatment for speech and language in primary progressive aphasia and primary progressive apraxia of speech: A systematic review. *Neuropsychology Review*.

Stephen M. Wilson, Maya L. Henry, Max Besbris, Jennifer M. Ogar, Nina F. Dronkers, William Jarrold, Bruce L. Miller, and Maria Luisa Gorno-Tempini. 2010. Connected speech production in three variants of primary progressive aphasia. *Brain*, 133(7):2069–2088.

Borui Yang, Md Afif Al Mamun, Jie M. Zhang, and Gias Uddin. 2025. Hallucination detection in large language models with metamorphic relations. *Preprint*, arXiv:2502.15844.

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):1–13.

# A Dataset statistics

The dataset and detailed methodology are described in (Rezaii et al., 2022), we provide a summary of the relevant details here.

All participants were shown a drawing of a family at a picnic from the Western Aphasia Battery-Revised (Clark et al., 2020) and were asked to describe it using as many full sentences as possible. The picture description task was administered by trained speech-language pathologists in a quiet room as part of a comprehensive clinical evaluation protocol. To prepare the written data, responses were recorded using an Olympus VN-702PC Voice Recorder, transcribed into text using the Microsoft Dictate application, and then manually verified for accuracy by a human expert who was blinded to the group assignments. Importantly, prosodic elements such as hesitations ("um," "uhh") and other disfluencies were carefully preserved in the transcripts, as these features are critical for capturing speech patterns characteristic of primary progressive aphasia. A total of 79 interviews with PPA patients were sourced from a study conducted within the PPA program at the Frontotemporal Disorders Unit of Massachusetts General Hospital (MGH). Expert neuropsychiatrists and speech-language pathologists carried out the assessment and annotation based on established clinical criteria. The dataset also includes 53 healthy controls, sourced from the

Speech and Feeding Disorders Laboratory at Massachusetts General Hospital (MGH) and Amazon's Mechanical Turk (MTurk). MTurk participants completed the short version of the Everyday Cognition test (12 items) to screen for cognitive decline and followed the same picture description protocol remotely. The distribution of subtypes is shown in Table 3 in Appendix A. All participants were native English speakers with no self-reported history of brain injury or speech/language disorders. Healthy controls and PPA patients were matched in terms of age, gender, handedness, and years of education.

| **Subtype** | **Nb. of Samples** |
|---|---|
| Logopenic Variant (lvPPA) | 26 |
| Semantic Variant (svPPA) | 24 |
| Nonfluent Variant (nfvPPA) | 29 |
| Healthy Controls | 53 |

Table 3: Distribution of subtypes and number of samples in the *original* version of the dataset.

| **Subtype** | **Nb. of Samples** |
|---|---|
| Logopenic Variant (lvPPA) | 433 |
| Semantic Variant (svPPA) | 402 |
| Nonfluent Variant (nfvPPA) | 335 |
| Healthy Controls | 960 |

Table 4: Distribution of subtypes and number of samples in the *expanded* version dataset.

| **Dataset** | **Mean** | **Median** | **Std. Dev.** |
|---|---|---|---|
| *Original version* | 132.98 | 104.00 | 89.47 |
| *Expanded version* | 7.76 | 7.00 | 4.93 |

Table 5: Statistics (mean, median, standard deviation) of text lengths (in words) for the *original* and *expanded* datasets.

## B Models

- **BERT**: A bidirectional transformer that captures context from both left and right of a word, making it effective for tasks that require deep semantic understanding (Devlin et al., 2019).

- **RoBERTa**: A robustly optimized version of BERT with improved training strategies and increased training data, designed to improve performance on a variety of NLP tasks (Liu et al., 2019).

- **MentalBERT**: A domain-specific transformer model fine-tuned on mental health-related text, aimed at capturing linguistic patterns specific to this domain (Ji et al., 2022).

- **ClinicalBERT**: A transformer fine-tuned on clinical text, optimized for healthcare-related tasks and well-suited for medical and diagnostic datasets (Alsentzer et al., 2019).

- **LLAMA:** `meta-llama/Meta-Llama-3-8B-Instruct`, sourced from the Hugging Face repository, developed by Meta, with 8 billion parameters, fine-tuned for instruction-based tasks (Touvron et al., 2023).

- **Mistral:** `mistralai/Mistral-7B-Instruct-v0.2`, sourced from the Hugging Face repository, developed by Mistral AI, with 7 billion parameters, optimized for instruction-based and conversational tasks (Jiang et al., 2023).

- **GPT-3.5-turbo:** Developed by OpenAI, a 175 billion parameter model known for its general-purpose conversational and reasoning capabilities (Brown et al., 2020).

- **GPT-4o-mini:** Developed by OpenAI, a lightweight variant of GPT-4, fine-tuned for optimized performance on smaller computational setups (OpenAI, 2023).

## C Prompt for Clinical Text Classification

The following prompt was used to guide the clinical text classification task performed by the LLMs:

You are a clinical text classifier specializing in language and speech characteristics related to Primary Progressive Aphasia (PPA). Based on the provided interview transcript of a patient, classify the text into one of the following categories:

- **lvPPA**: Logopenic Variant, Characterized by word-finding difficulties and impaired repetition abilities. Patients may frequently pause or hesitate as they search for words, and they may struggle to repeat phrases accurately.
Example: Patient might say, "I went to the... um... place where... you know, people get... books," when trying to say "library." They may also struggle to repeat phrases accurately, often omitting words or stumbling.

- **svPPA**: Semantic Variant, Primarily affects the understanding of word meanings (semantic knowledge). Patients may struggle with naming and comprehension, even for common objects. They often resort to broad categories instead of precise words (e.g., thing instead of fork).
Example: When shown a picture of a dog, the patient might say, “It’s an animal… I think it’s a pet,” without being able to retrieve the word "dog." They may also have difficulty understanding specific terms, relying on broader descriptions.

- **nfvPPA**: Impacts grammar and speech production, leading to slow, effortful, and agrammatic speech. Patients may omit small grammatical words (e.g., “is,” “the”) and speak in a telegraphic manner. Patients tend to use very short sentences, a rich vocabulary with low-frequency words, and more nouns compared to verbs.
Example: The patient might say, “Walk… store… buy milk,” instead of “I’m going to walk to the store to buy milk.” Speech is often halting and labor-intensive, with noticeable pauses.

- **control**: The individual demonstrates fluent, grammatically correct speech, free from any markers of hesitation, effortful speech, or semantic impairment. There are no indications of word-finding difficulties or grammatical errors. The individual uses both simple and complex sentences naturally and appropriately. They can express themselves clearly without notable pauses, hesitations, or substitutions. The vocabulary used is appropriate for the context, and their language comprehension and responses are cohesive.
Example: “I’m going to walk to the store to buy some milk” or “After I finish work, I plan to go for a walk and then cook dinner.” The language is fluent, natural, and demonstrates coherent sentence-building abilities.

Analyze the language, sentence structure, vocabulary, and speech flow within the conversational context of the interview to determine the most fitting category. Your response should include only one of the following labels: **lvPPA**, **svPPA**, **nfvPPA**, or **control**. If the text does not clearly fit into one category, analyze it carefully and suggest the most likely category based on available evidence.

## D Computational Resources

The experiments described in Section 4.2 and 6.2 were conducted on Google Colab Pro using an NVIDIA L4 GPU.
The experiments described in Section 5 were conducted using two different computational setups. For LLAMA and Mistral, we ran experiments locally on a system running Ubuntu 22.04.4 LTS (Jammy Jellyfish). This system featured an AMD Ryzen 9 7950X 16-Core Processor (32 threads, 16 cores, 2 threads per core) with a maximum clock speed of 5.88 GHz, 62 GB of RAM, 2 GB of swap space, and an NVIDIA RTX A6000 GPU with 48 GB of memory, using CUDA 12.4 for GPU acceleration. For GPT-3.5 and GPT-4o-mini, we relied on the OpenAI API, accessing the models via cloud-based inference.

## E Results of Traditional Machine Learning’s experiments

This section presents the results of the remaining traditional ML experiments conducted in our study. For each classification model, we include performance metric plots across the five cross-validation folds. These graphs offer a more comprehensive view of model behavior and complement the summary statistics discussed in the main text.

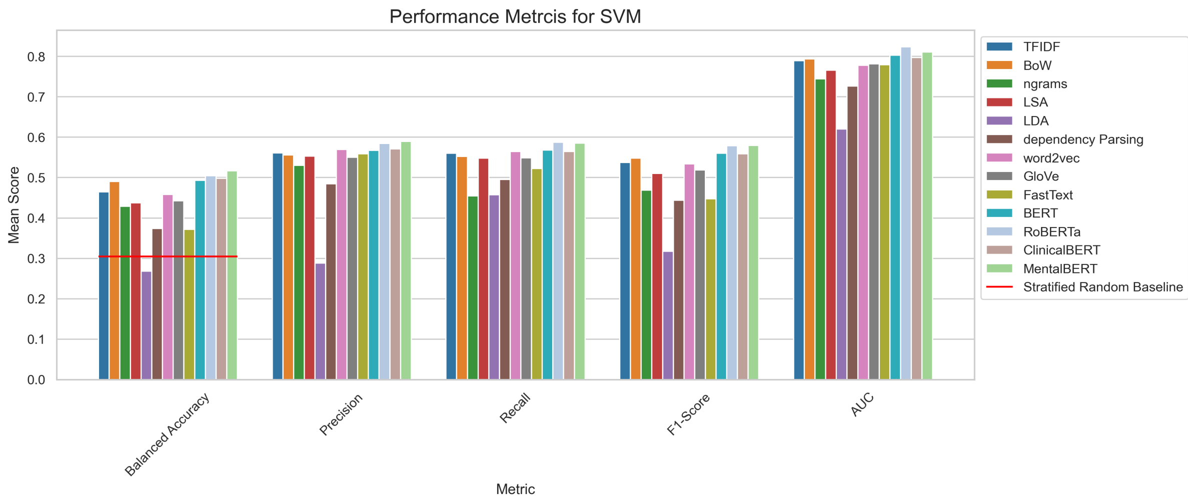


Figure 4: SVM performance

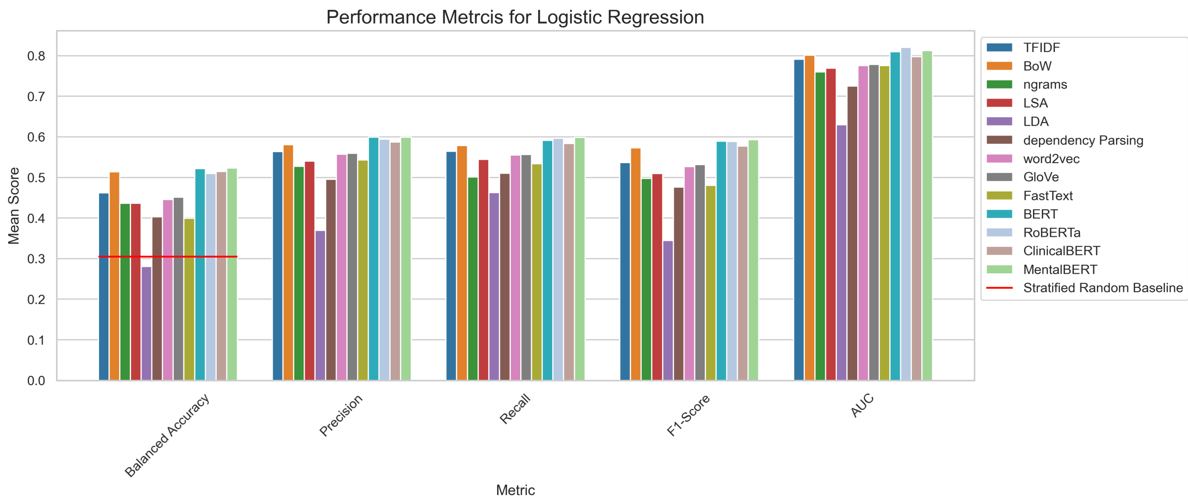


Figure 5: Logistic Regression performance

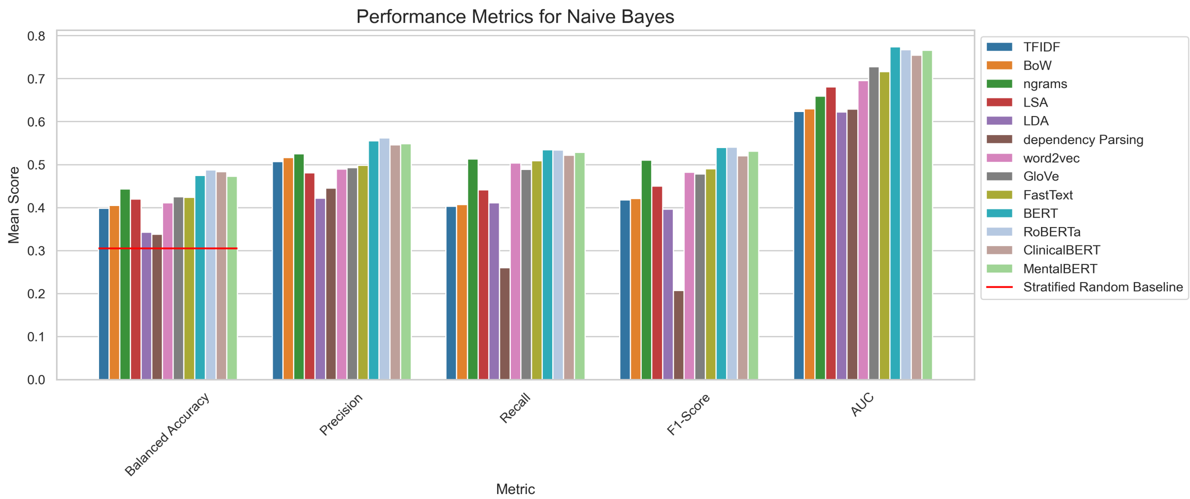


Figure 6: Naive Bayes performance

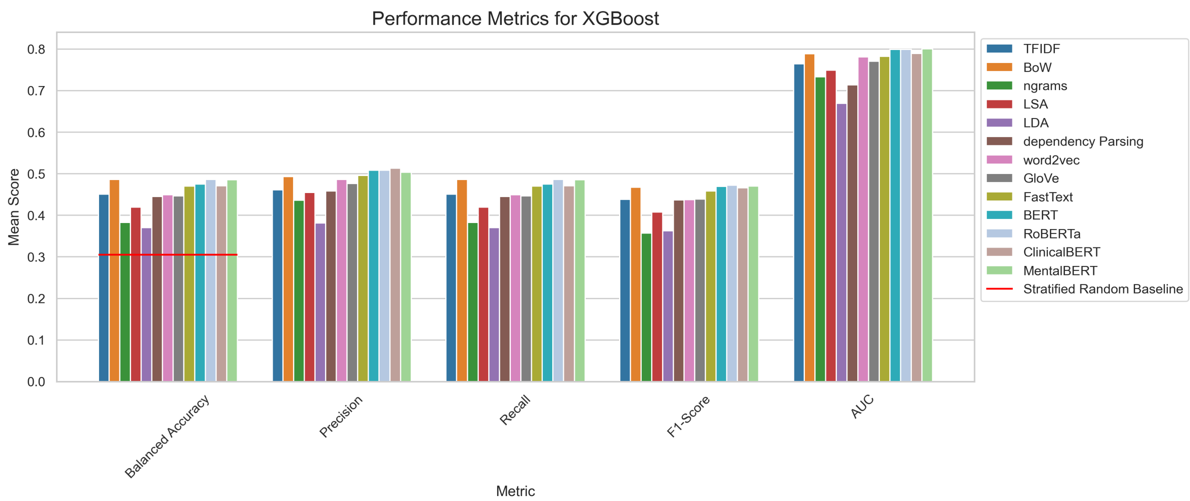


Figure 7: XGBoost performance

# F Results of LLMs' Experiments

This section presents the performance of LLMs. We report key metrics such as balanced accuracy, precision, recall, and F1-score across all models. Results are visualized using bar charts for comparative clarity. Additionally, confusion matrices are provided to highlight subtype-specific strengths and weaknesses, offering a more granular view of the classification outcomes.

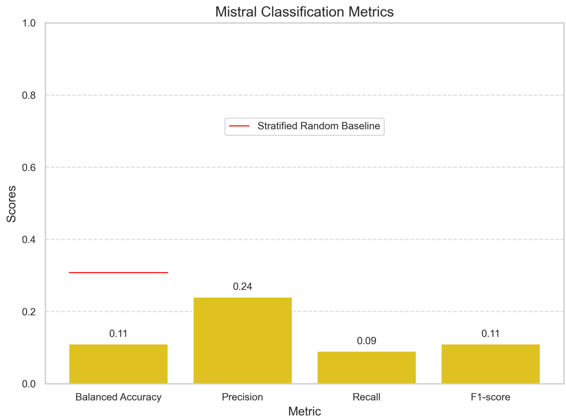


Figure 8: Mistral Performance

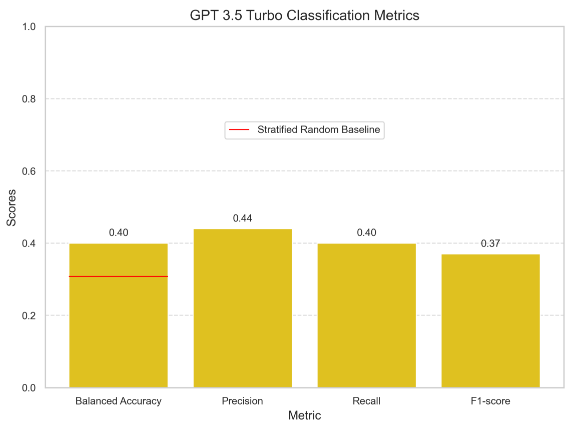


Figure 9: GPT-3.5 Performance

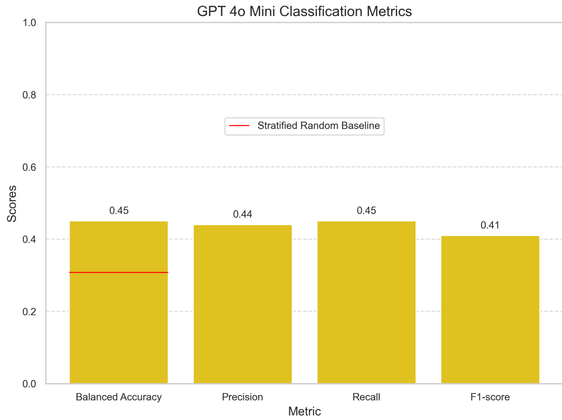


Figure 10: GPT-4o-minia performance

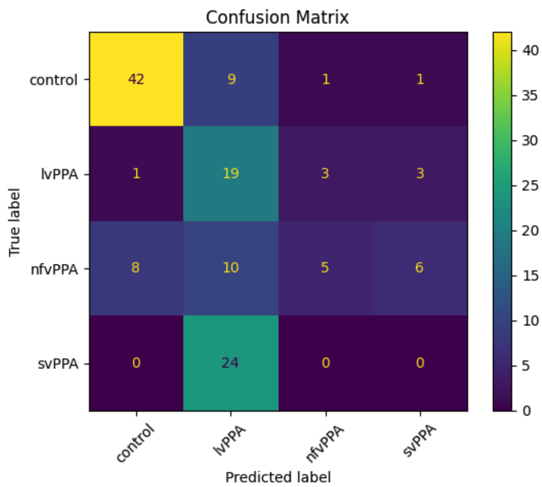


Figure 11: LLAMA Confusion Matrix

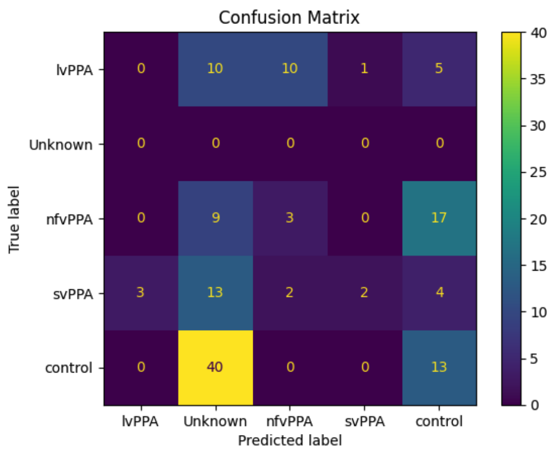


Figure 12: Mistral Confusion Matrix

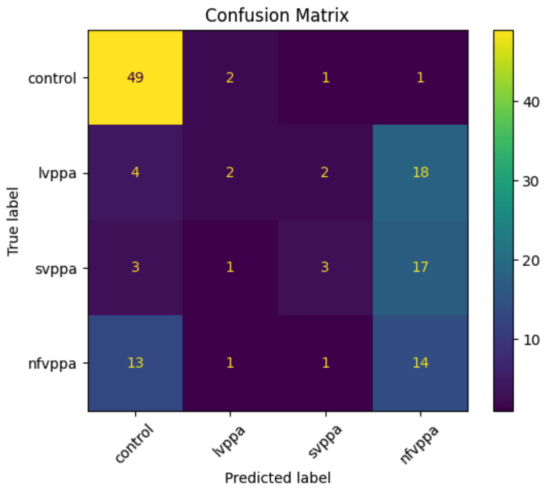


Figure 13: GPT-3.5-Turbo Confusion Matrix

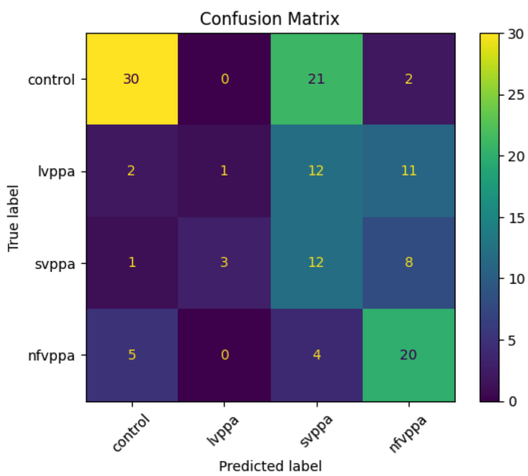


Figure 14: GPT-4o-mini Confusion Matrix

# G Feature Importance Analysis with LIME

To better understand model behavior and interpret classification decisions, we conducted a feature importance analysis using the LIME framework. This approach allows us to identify which input features most influenced individual predictions, providing insights into the linguistic patterns leveraged by the models for each PPA subtype.

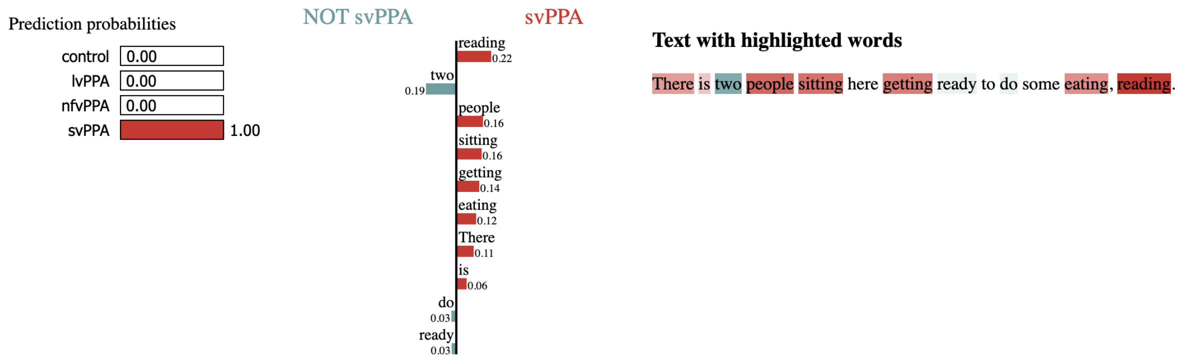


Figure 15: Token-level feature importance estimated by LIME for a svPPA representative sample - MLP + MentalBERT's features.

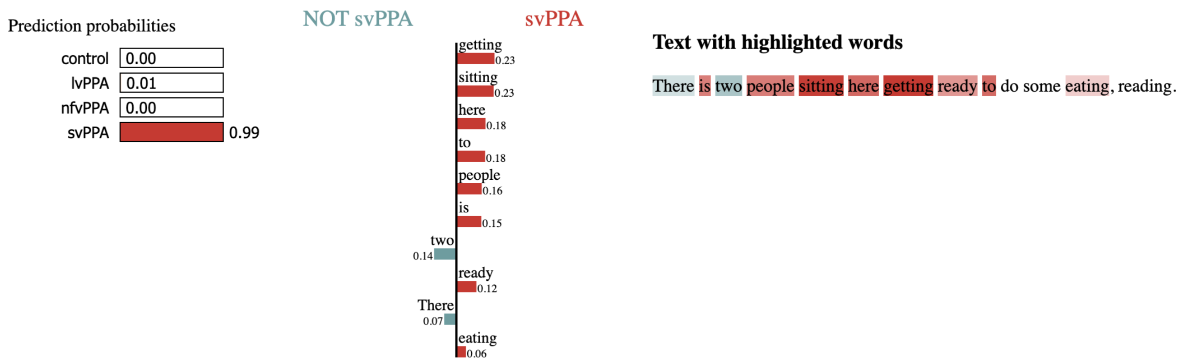


Figure 16: Token-level feature importance estimated by LIME for a svPPA representative sample - BERT.

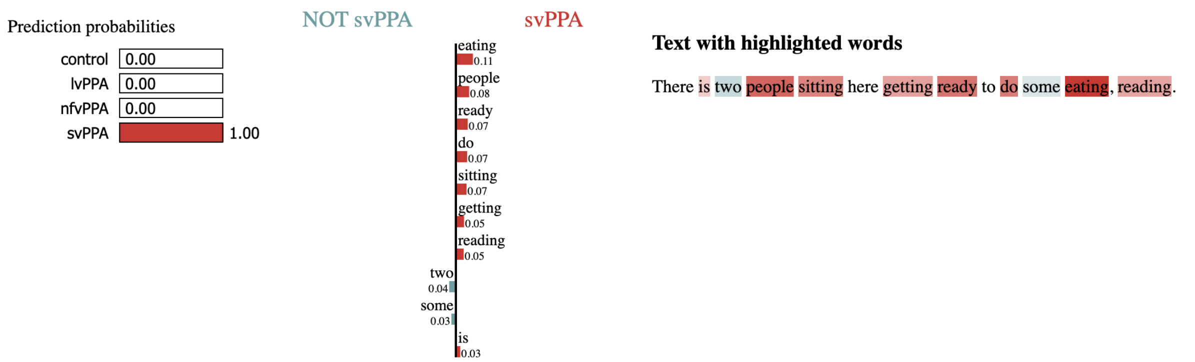


Figure 17: Token-level feature importance estimated by LIME for a svPPA representative sample - RoBERTa.

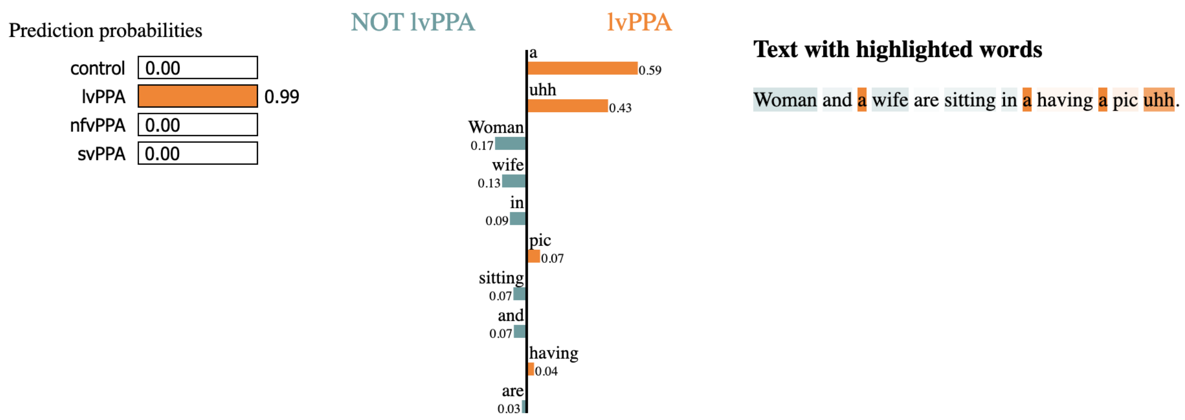


Figure 18: Token-level feature importance estimated by LIME for a lvPPA representative sample - MLP + Mental-BERT's features.

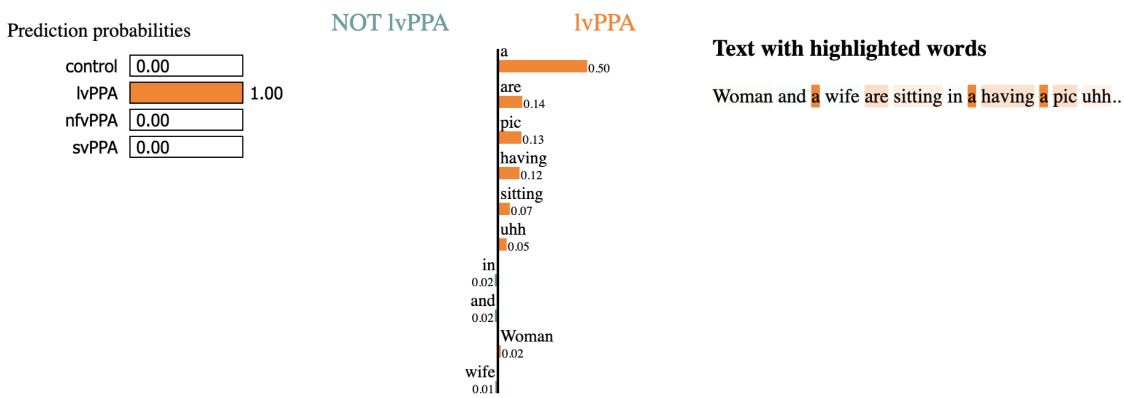


Figure 19: Token-level feature importance estimated by LIME for a lvPPA representative sample - BERT.

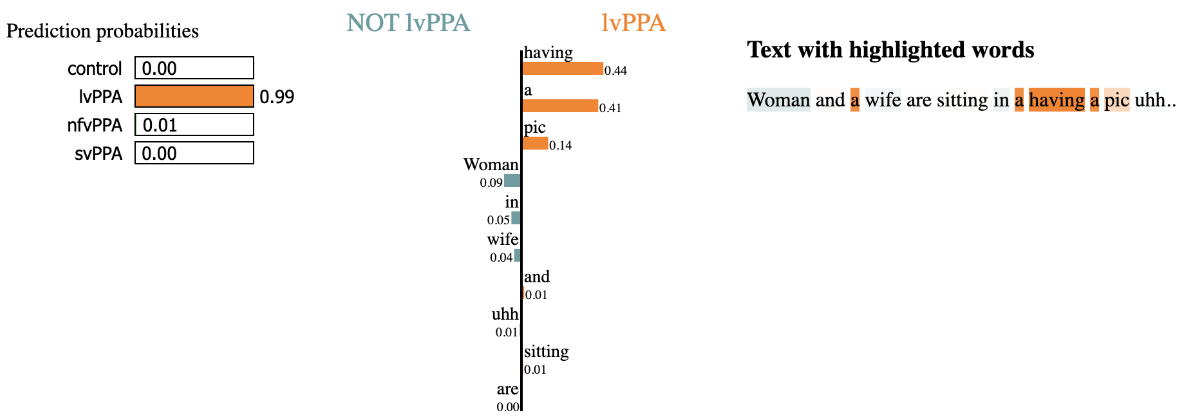


Figure 20: Token-level feature importance estimated by LIME for a lvPPA representative sample - RoBERTa.

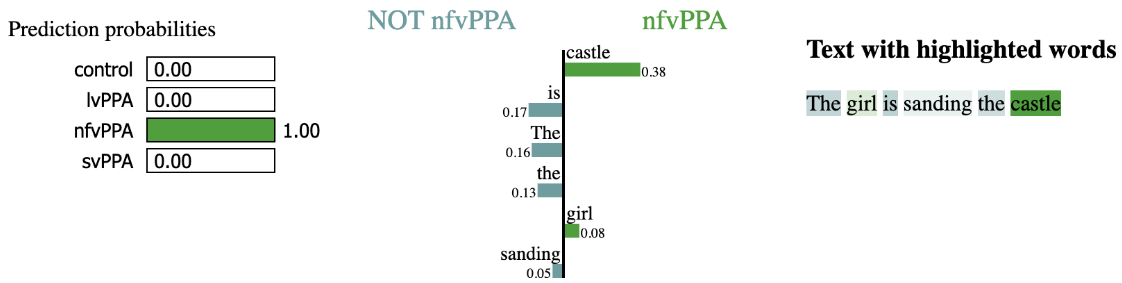


Figure 21: Token-level feature importance estimated by LIME for a nfvPPA representative sample - MLP + MentalBERT's features.

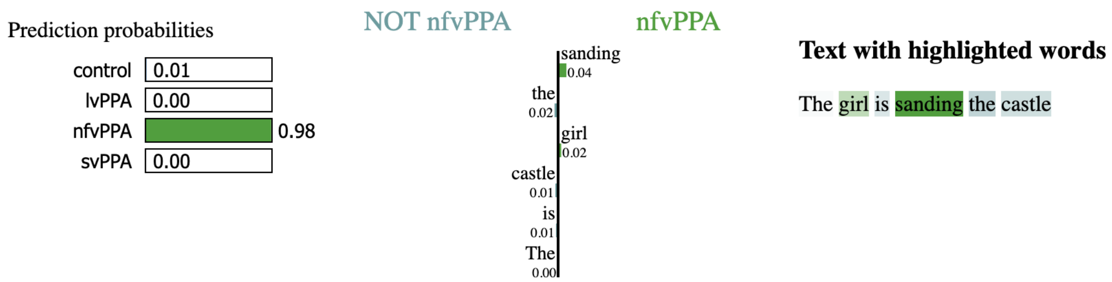


Figure 22: Token-level feature importance estimated by LIME for a nfvPPA representative sample - BERT.

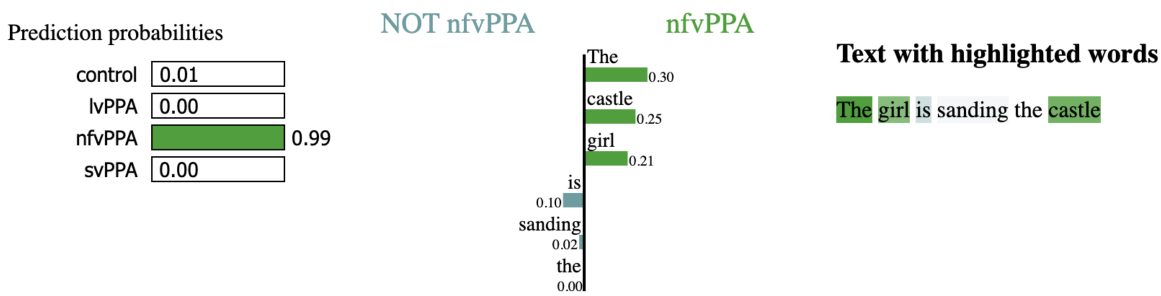


Figure 23: Token-level feature importance estimated by LIME for a nfvPPA representative sample - RoBERTa.