Determination of Profiles between the Autism Disorder and ADHD Disorder for Diagnostic Support using Large Language Models

Monica Hernandez Lordui
mhernandezlo@umass.edu

1 Problem statement

Given that Autism Spectrum Disorder (ASD) is now considered a type of epidemic (Newschaffer et al., 2007), it is even causing the United States government to consider it a problem of statistical relevance (U.S. Department of Health and Human Services, 2025). Individuals with autism often receive misdiagnoses (including inappropriate pharmacological and therapeutic treatments), which leads to poor quality of life. This problem high lights the challenge of differentiating the inherent characteristics of autism from its comorbidities and overlapping features (May et al., 2021), (Fusar-Poli et al., 2022). In both disorders, regarding the attentional symptom, the individual can only focus intensely on their special interest (ASD), but is incapable of sustaining attention on everyday tasks (ADHD). For example, they may be highly productive writing algorithms, but could return several hours later and notice they left their car door wide open or leaving a heating element of their stove on without noticing until the next day. Due to this comorbidity, they tend to have many accidents or high-risk situations. The issue of comorbidities between Autism Spectrum Disorder (ASD) and Attention-Deficit/Hyperactivity Disorder (ADHD) is a crucial and very common topic in neuropsychiatric diagnosis. Previously, diagnostic manuals (DSM-IV) prevented an individual from being diagnosed with both conditions simultaneously. However, the DSM-5 (2013) eliminated that restriction, officially recognizing that ASD and ADHD frequently coexist in the same individual. The coexistence of ASD and ADHD is the most common comorbidity in neurodevelopment. It is estimated that between 50% and 70% of children diagnosed with ASD also meet the criteria for ADHD. Approximately 20% to 50% of individuals with ADHD also meet the criteria for ASD (Simonoff et al., 2008), (Matson et al., 2013). This project aims to use the advantages of Large Language Models (LLMs) to establish differentiating profiles within the autism spectrum and ADHD disorder, in a sense, shape the overlap. The project objectives are the following:

- Leverage the capabilities of Large Language Models (LLMs) to enhance the diagnostic specificity of Autism Spectrum Disorder (ASD), particularly focusing on differentiating profiles within the spectrum and quantifying the symptomatic overlap with comorbid conditions.

- Quantify the significant overlap between conditions, going beyond simple coexistence. Specifically, I seek to determine the likelihood that an autistic individual's symptom set aligns more closely with one disorder or comorbidity than another.

- Determine Symptom Alignment Likelihood: Specifically, to assess the probability that an autistic individual's symptom profile aligns more closely with one specific disorder or comorbidity than another.

- Support Clinical Decision-Making: To serve as a powerful clinical decision-support tool for practitioners, providing immediate access to current research and focused treatment strategies, especially for complex, multilayered conditions like ASD.

2 What you proposed vs. what you accomplished

The following objectives were proposed and all were fulfilled:

- Collect and preprocess dataset. Augmented Data Generation and Pre-processing using

https://github.com/mhernandezlordui/Final-Project-LLM

the Few-Shot Learning strategy via the Gemini API. The data's tone and format are defined.

- Build and train of ClinicalBERT was on collected dataset and examine its performance. The process consisted of three main phases executed sequentially: dataset preparation, model initialization and optimization, and training and evaluation loop.
- Creation of the code that implements the clinical differential diagnosis to compare a new patient narrative with the ideal vector profiles for Autism Spectrum Disorder (ASD) and Attention Deficit Hyperactivity Disorder (ADHD).
- Error analysis when changing the reference values ”ground truth” narratives.

3 Related work

Autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD) are increasing in prevalence and often occur together (co-occur). Work on comorbidity between ASD and ADHD can be found in (Mayes et al., 2012). The study aimed to determine the degree to which core ADHD and autistic symptoms overlap in and discriminate between children aged 2 to 16 with autism and ADHD.

The study found that all children with ASD presented 15 or more of the 30 symptoms on the Checklist for Autism Spectrum Disorder (CASD), while none of the children with ADHD did. Another conclusion of this work was that 30 autistic symptoms were found in more than half of the children with ASD, but none were present in the majority of children with ADHD (both inattentive and combined types). Additionally, they found evidence to indicate that ADHD symptoms are common in autism and that autistic symptoms are not common in ADHD. From the above, they were able to point out three elements: The core symptoms of ADHD (attention deficit, impulsivity, and hyperactivity) are considered part of autism. Children with low- and high-functioning ASD, and children with Combined-type ADHD, did not differ on maternal ratings of attention deficit, impulsivity, and hyperactivity. Due to the high frequency of ADHD symptoms in autism, children with ASD may initially be misdiagnosed with ADHD.

In the work of (Antshel and Russo, 2019), the focus was on the differential diagnosis of the comorbid state (ASD+ADHD). Among the findings, they discovered that ASD and ADHD have shared genetic heritability and are both associated with shared deficits in social functioning and executive functioning. However, there are quantitative and qualitative differences in the phenotypic presentations of the deficits that characterize ASD and ADHD. For interventions for ASD to be as effective as possible, comorbid ADHD must be considered (and vice versa). The general conclusion in this article stated that ASD and ADHD suggest some overlap between the two disorders, but also sufficient differences to indicate that these conditions are distinct enough to warrant separate diagnostic categories.

In the work of (Taurines et al., 2012) the same conclusion as (Mayes et al., 2012) was reached: a high rate of ADHD prevailed in autism and autism spectrum disorders (ASD). Due to several similar studies that presented this conclusion, in the planned revision of the DSM-IV TR, the DSM-5, the diagnoses of autistic disorder and ADHD will no longer be mutually exclusive. The work presented a review of data on comorbidity rates and symptom overlap, and discusses common and disorder-specific risk factors, including recent proteomic studies. They compared areas of great comorbidity and overlap, such as, areas of attention, reward processing, and social cognition.

Addressing the problem of comorbidity and the diagnosis of ADHD and ASD disorders itself is relatively new. No applications of LLMs (Large Language Models) to the comorbidity (ASD + ADHD) were found. However, few and recent papers applying LLMs to diagnose or support decision-making for autism were found.

In (Mukherjee et al., 2023), ASD detection in a child was implemented using parent dialogue. The BERT model and ChatGPT were used to analyze and detect ASD symptoms. They applied sentiment analysis techniques to determine whether a sentence expressed positive or negative feelings regarding ASD symptoms. The system selected only the sentences with a positive sentiment for further analysis using the cosine similarity model. Other low-citation works due to novelty are found in (Chien and Tai, 2024) and (Ren et al., 2023).

https://github.com/mhernandezlordui/Final-Project-LLM

4 Your dataset

The dataset was constructed in 3 blocks:

1. Block: Few-Shot Strategy definition. The objective of this step was to define the tone and format of the data. To achieve this, a long text string (a constant) was defined containing 10 examples of carefully labeled clinical narratives (pure ASD, pure ADHD, Comorbidity, and Control). The construction of these examples was done using official diagnostic tests, and these same tests were used to obtain the labels for Asperger-type ASD and ADHD (Data Annotation).The desired outcome was to establish a ”patient profile” that the LLM would use as a guide to imitate the tone, non-clinical language, and the way symptoms should be intertwined within the text.The official tests used were: DIVA-5 (ADHD), ASRS-5 (ADHD), ADHD (Inattentive Type) DSM-5, EQ (ASD), AQ-10 (ASD), ASDI (ASD), The Ritvo Autism Asperger Diagnostic Scale-Revised (RAADS-R), and AAA- Adult Asperger Assessment.

2. Block: Narratives Generation. The prompt was constructed with the desired guidelines, and the API was called for the content generation task. It was requested to create a single, coherent narrative that contained the specified symptoms. These symptoms were randomly selected (between 3 and 30). This random selection was possible from the total pool of 36 criteria. The value of this random selection affected the precision of the symptom classification model, generating overfitting or making the training epochs insufficient. The training in the classification and fine-tuning task that achieved good performance was the one already mentioned. The criteria themselves were obtained using the DSM-5 for Asperger-type ASD and Inattentive-type ADHD disorders.The Few-Shot Prompting technique was used. This is a Prompt Engineering technique employed to improve the quality and accuracy of the LLM’s response for a specific task. It consists of providing the LLM with a task description and a small number of question-answer pairs directly within the prompt (the instruction). In text generation, this is used to imitate the desired style and reasoning for the task without the need for formal fine-tuning of the internal weights. The goal was to generate and imitate the tone of the clinical narrative, limited by the questions from the official tests, which are conversational in nature—that is, very simple language, yet driven by the deficiencies of the disorder.

3. Block: Collection, Labeling, and Batch Saving Process. A .csv file was created and saved in batches due to the complexity of the task and memory requirements. The header included the ”Text” column (for the generated narratives) and 36 label columns: 18 labels for Asperger ASD symptoms and 18 for Predominantly Inattentive ADHD.These two subtypes were chosen because they represent the highest comorbidity. The criteria were extracted from the DSM-5 sections under the codes: 299.00 (F84.0), 315.4 (F82), 314.01 (F90.0, F90.1, F90.2), and 314.00 (F90.0), 307.3 (F98.4), 307.23 (F95.2), and 307.22 (F95.1), 307.21 (F95.0). A example in the dataset was: Narrative generate: ”My life feels like a constant scramble; I can never seem to get my tasks organized, and I’m always misplacing my keys or important papers, which makes everything worse. It’s not that I don’t try to listen, but when people talk to me directly, my mind just drifts off, making me seem like I’m not paying attention and struggling to sustain focus. This often leads to me blurting things out or cutting people off mid-sentence without meaning to, which I know can be annoying. Social situations are a minefield; I struggle to understand what’s expected of me or how to act, and I’ve never really been good at making friends or connecting with others my age. I just don’t get the subtle cues like gestures or facial expressions. Sometimes, when I’m really overwhelmed, I’ll find myself flapping my hands or rocking back and forth, and I have this odd fascination with spinning the wheels on toy cars for ages. My days have to follow a strict routine; even a small change to my schedule sends me completely off kilter. I’d rather spend hours delving into my obsession with historical train models than try to navigate a social gathering, where I often feel the urge to just get up and walk away mid-conversation...cut here”

Generating multi-label narrative (15 symptoms) with:

$$[{}'ADHD\_Organization','ASD\_Motor\_Rep', 'ASD\_Relationships','ASD\_Verbal\_Rigid', 'ADHD\_Interrupts','ASD\_Context\_Adjust', 'ADHD\_Leaves\_Seat','ADHD\_Not\_Listen', 'ADHD\_Loses\_Things','ASD\_Preoccupation', 'ASD\_Comp\_Gestures','ASD\_Transition', 'ASD\_Peers','ADHD\_Sustain\_Attn', 'ASD\_Obj\_Rep']$$

https://github.com/mhernandezlordui/Final-Project-LLM

In the CSV file, the narrative is assigned to the 'text' field, and the final binary vector (the Ground Truth) is constructed with the narrative text. 1000 samples were generated using the 10 initial samples obtained from the official tests. The embedding space was demarcated by: pure ASD-Asperger DSM-5, pure ADHD (inattentive type) DSM-5, DIVA-5 (ADHD), ASRS-5 (ADHD), EQ (ASD), AQ-10 (ASD), ASDI (ASD), 1 example based in ASD case, Comorbidity (ASD + ADHD) 20% ASDI, 10% EQ, 10% AQ-10 and 20% DSM-5, comorbidity (ASD + ADHD) . 70% ADHD, 30% ASD-Asperger.

The list of tests can be found in: (NIC), (Baron-Cohen et al., 2006), (NIC, 2012), (Gillberg et al., 2001), (Kooij and Francken, 2019), (American Psychiatric Association, 2013), and (Ritvo et al., 2011).

4.1 Data annotation

The annotation task in this project is directly related to the construction of the 36-label binary vector for each generated narrative. This is not a traditional annotation but a programmatic self-annotation controlled by the code. The sampling algorithm $random.sample$ acts as the annotator, used in Block 3. The code decides which labels must be present. It annotates the presence 1 or absence 0 of each of the 36 symptoms for the current narrative. the variable narrative_text was generated by the LLM with the explicit instruction to include only the symptoms selected in the variable $sampled\_labels\_names$. Therefore, the programmatically generated binary vector is the narrative of symptoms that the ClinicalBERT model must learn.

Extensions of the reference values were made in order to integrate the 36 symptoms into pure, short, and colloquial syntax. Pure and mixed results were tested. But unfortunately, the highest precision and probability were achieved with the smallest set shown in Table No. 1. The results will show how the comorbidity diagnosis of the symptoms is modified when this value is changed. falta hacer en seccion de recultados

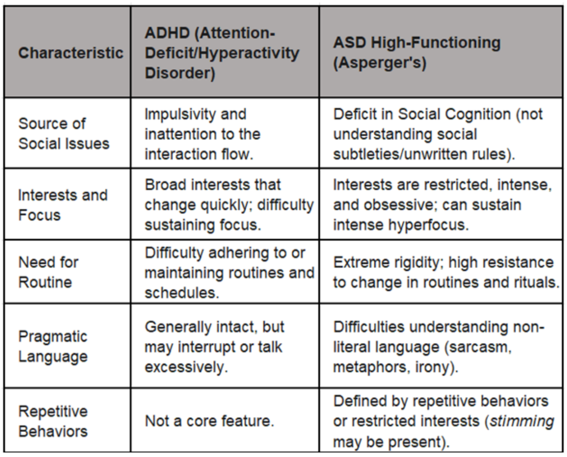
5 Baselines

The Reference Values are understood as the Average Vectors $(reference\_vector\_asd, reference\_vector\_adhd)$ in the code. These were calculated using Table No. 1 shown. These vectors represent the ideal semantic core in the vector space. These are created by utilizing the body of the ClinicalBERT model (previously fine-tuned) to extract and average the embeddings of narratives that describe solely the fundamental and non-overlapping criteria for each disorder accord DSM-5. These vectors represent the ideal semantic core of ASD and ADHD in the vector space. The diagnostic function uses these vectors as fixed 'focal points.' Upon entering a new patient narrative, its embedding is compared with these two focal points using Cosine Similarity, and the results are normalized with Softmax to produce the final Percentage Probabilities (ASD_Porcentaje, DHD_Porcentaje, indicating the patient's vectorial proximity to each diagnosis.

The final choice of ClinicalBERT particularly Bio_ClinicalBERT was due to a performance analysis that disqualified the other two options based on critical failures during the validation phase of the symptom classification task. Particularly, BERT exhibited overfitting. This means it learned the 1000 training samples almost perfectly (with a very low Training Loss), but failed dramatically when attempting to generalize to the new narratives in the validation set. This indicates that it memorized the synthetic dataset instead of learning the underlying patterns.

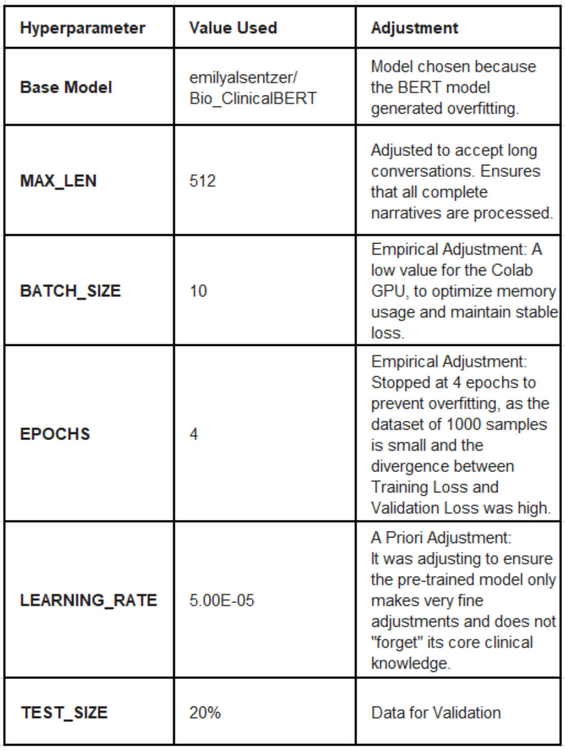
The classification Model used for fine-tuning works by classifying a text narrative into a 36-label binary vector. This is a multi-label classification problem. This approach was chosen due to the very definition of comorbidity, which is what the project aims to model. The ClinicalBERT body processes the sequence. The final layer $(AutoModelForSequenceClassification)$ takes the output $(PoolerOutput)$ and passes it through a linear layer with 36 output neurons. These outputs (logits) are combined with a Sigmoid function to produce probabilities, which are then used for Binary Cross-Entropy (BCE) to calculate the error. The model that performs the diagnosis is a vector model residing in the embedding space. For this reason, the best fine-tuned and validated model is saved along with its weights, to be used as a tokenizer (or embedding extractor) in order to process the inputs from a consulting user and the pure criteria shown in Table No. 1

Table No. 1



Characteristic	ADHD (Attention-Deficit/Hyperactivity Disorder)	ASD High-Functioning (Asperger's)
Source of Social Issues	Impulsivity and inattention to the interaction flow.	Deficit in Social Cognition (not understanding social subtleties/unwritten rules).
Interests and Focus	Broad interests that change quickly; difficulty sustaining focus.	Interests are restricted, intense, and obsessive; can sustain intense hyperfocus.
Need for Routine	Difficulty adhering to or maintaining routines and schedules.	Extreme rigidity; high resistance to change in routines and rituals.
Pragmatic Language	Generally intact, but may interrupt or talk excessively.	Difficulties understanding non-literal language (sarcasm, metaphors, irony).
Repetitive Behaviors	Not a core feature.	Defined by repetitive behaviors or restricted interests (*stimming* may be present).

The following is a list of the hyperparameter:



Hyperparameter	Value Used	Adjustment
Base Model	emilyalsentzer/ Bio_ClinicalBERT	Model chosen because the BERT model generated overfitting.
MAX_LEN	512	Adjusted to accept long conversations. Ensures that all complete narratives are processed.
BATCH_SIZE	10	Empirical Adjustment: A low value for the Colab GPU, to optimize memory usage and maintain stable loss.
EPOCHS	4	Empirical Adjustment: Stopped at 4 epochs to prevent overfitting, as the dataset of 1000 samples is small and the divergence between Training Loss and Validation Loss was high.
LEARNING_RATE	5.00E-05	A Priori Adjustment: It was adjusting to ensure the pre-trained model only makes very fine adjustments and does not "forget" its core clinical knowledge.
TEST_SIZE	20%	Data for Validation

6 Your approach

Tools used. For language models and Deep Learning, the PyTorch library was used, specifically the torch and transformers modules. To generate the training data (data augmentation), the Client library from google.genai was imported. Scikit-learn was part of the training process and metrics calculation, Scipy was used to calculate the cosine similarity, and Numpy and Pandas were used for handling data and numerical calculations.

Problems in the implementation I encountered a large number of problems when starting out. While testing my code for generating the synthetic clinical narratives from users, Google kept stopping the execution, telling me I had exceeded the number of daily requests—it was supposed to be 20 per day—and that I had to wait a few minutes. Then I would wait, and it would happen again. Therefore, I had to pay for an exclusive Gemini client API from the google.genai library for that generation process. After making that change, I was able to generate the augmented data without any problem or unexpected stops.I also opted to pay 11 dollars to eliminate the runtime disconnections in Colab, which were driving me crazy and halting processes that took one or two hours; it was terrible. Additionally, I disabled the hibernation permissions on my PC because the augmented data generation of 1000 lines took 3 hours and 20 minutes.

Methodology To solve the problem presented, the methodology shown in Figures 1 and 2 was utilized. The goal was to train a Language Model (LLM) to classify the previously defined symptoms of two disorders: ASD and ADHD. The LLM had to rely solely on the conversation with the user. The project did not possess clinical conversations to guide the language model, as this data is confidential. The project also lacked a 'conversation model' of a user who suspects having high-functioning autism or ADHD. However, the existence of official autism and ADHD tests was known. These tests consist of a group of colloquial, easy-to-understand questions for the patient, the syntax of which semantically encapsulates the disorder's symptoms. For this reason, the idea arose to use the official autism/ADHD tests to generate realistic conversations. Consultations with the Gemini assistant suggested that between 1000 and 5000 clinical data points were required for reliable language model training. Unfortunately, only 1000 data points were used. Following the training and validation of the model, it would be utilized to generate or map the clinical vector representations (embeddings) of any user conversational input onto a symptom domain space specific to ASD (Asperger) and ADHD.

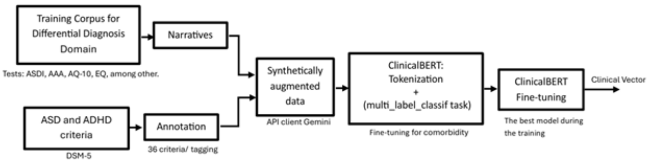

Training Corpus for Differential Diagnosis Domain
Tests: ASDI, AAA, AQ-10, EQ, among other.
Narratives
ASD and ADHD criteria
DSM-5
Annotation
36 criteria/ tagging
Synthetically augmented data
API client Gemini
ClinicalBERT: Tokenization + (multi_label_classif task)
Fine-tuning for comorbidity
ClinicalBERT Fine-tuning
The best model during the training
Clinical Vector

Fig. 1.

In Fig. 1., each box represents the steps required to obtain the model's fine-tuning. Initially, the most important diagnostic tests for ASD/ADHD are compiled. With Gemini's assistance and by observing the DSM-5 criteria, the 36 most representative criteria or symptoms of both disorders are obtained (Annotation box). For class balancing, 18 were obtained for ASD and 18 for ADHD. Also, with Gemini's help, the official test questions were converted into 10 basic narratives (Narratives box) with precise comorbidity instructions that were already explained in Section 4. The idea behind this was to configure the domain vector space for both disorders (Augmented Data box). Using these inputs, the ClinicalBERT model is trained for the multi-label classification task and then fine-tuned. The language model is thus trained to produce a clinical vector that represents a mathematical mapping of a symptom narrative. The code that resolve this part are gene_data.ipynb and training.ipynb

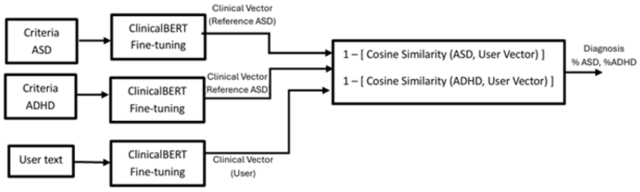


Fig. 2.

In Fig. 2. is shown the comorbidity calculation at the embedding level. The best-trained model is saved as a .pt instance; it was named clinicalbert_final_epoch_model.pt. For the calculation of the pure clinical reference vector is used the language model without the classification layer. Then, these vectors are the input arguments to the function that calculates the similarity cosine between the reference vector and the symptom vector. This is done for each disorder. Finally, the difference yields the comorbidity of the disorders, given the conversation of the user consulting the LLM. This process is illustrated in Fig. 2. and it is implemented by the function inferences.ipynb.

Functional implementation Functionally, the project consists of three Python scripts: *gene_data.ipynb*, *training.ipynb*, and *inferences.ipynb*. The first obtains augmented data with a conversational narrative, incorporating symptoms of both disorders. Technical and time constraints prevented the generation of more than 1,000 data points. Crashes on the running PC were still occurring. The *training.ipynb* script performs the classification and fine-tuning. It recognizes symptoms in a conversation with a user.

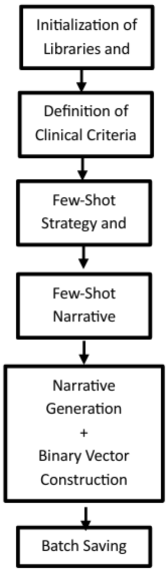

Initialization of Libraries and
Definition of Clinical Criteria
Few-Shot Strategy and
Few-Shot Narrative
Narrative Generation + Binary Vector Construction
Batch Saving

Fig. 3.

Finally, *inferences.ipynb* calculates the comorbidity of the disorders. In the Fig. 3. is shown the code flow of *gene_data.ipynb*. The first step is loading the pre-trained ClinicalBERT model. It automatically adds a linear classification layer with 36 output neurons, which is crucial for the multi-label task. Initially, the plain BERT model was tested, but it did not yield good results because it is a general-purpose model; the algorithm exploded. Next, the model is transferred to the GPU (previously configured in the Colab environment), the AdamW optimizer is defined, and the BCEWithLogitsLoss function is chosen, which was researched to work well with the multi-label classifier. Then, the model under training is evaluated using the F1-score and AUC-ROC. The performance is assessed per epochs, and the best-performing model is saved.

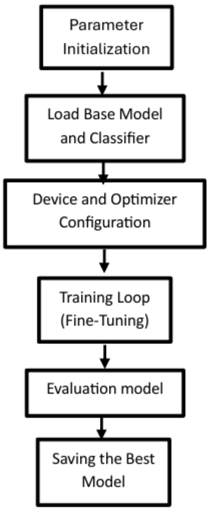

Parameter Initialization
Load Base Model and Classifier
Device and Optimizer Configuration
Training Loop (Fine-Tuning)
Evaluation model
Saving the Best Model

Fig. 4.

https://github.com/mhernandezlordui/Final-Project-LLM

In Fig. 5. are represented the steps in the script inferences.ipynb. Basically, the process begins by transforming the ClinicalBERT language model, previously fine-tuned on a multi-label classification task, into an efficient embedding extractor. This is achieved by removing the 36-neuron linear classification layer and retaining only the Transformer encoder body. This extractor becomes a tool capable of mapping any patient symptom narrative to a single 768-dimensional clinical vector. To establish the basis for differential diagnosis, two fixed reference vectors were generated: one for Autism Spectrum Disorder (ASD) and another for Attention Deficit Hyperactivity Disorder (ADHD). These reference vectors were obtained by averaging the embeddings generated from conceptually pure narratives of each disorder, thus establishing two immovable ”focal points” in the vector space. Once the space is configured, the differential diagnosis is performed vectorially. When a patient’s narrative is entered, the system converts it into a user vector and then measures its proximity to the fixed focal points of ASD and ADHD using Cosine Similarity. The resulting similarity scores are processed through the Softmax function, which normalizes them into differentiated percentage probabilities. The final output provides the probability that the user’s narrative aligns more semantically with the core symptoms of ASD or with the core symptoms of ADHD, thus providing a primary diagnosis based on vector proximity.

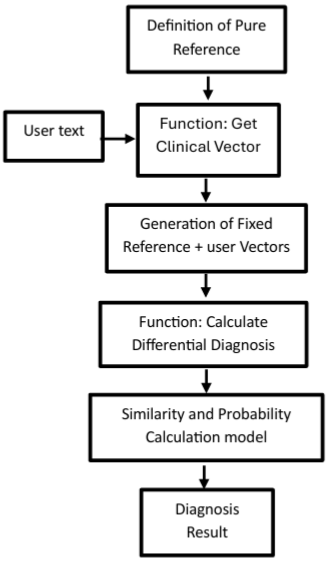

Definition of Pure Reference
User text
Function: Get Clinical Vector
Generation of Fixed Reference + user Vectors
Function: Calculate Differential Diagnosis
Similarity and Probability Calculation model
Diagnosis Result

Fig. 5.

7 Analysis and Results

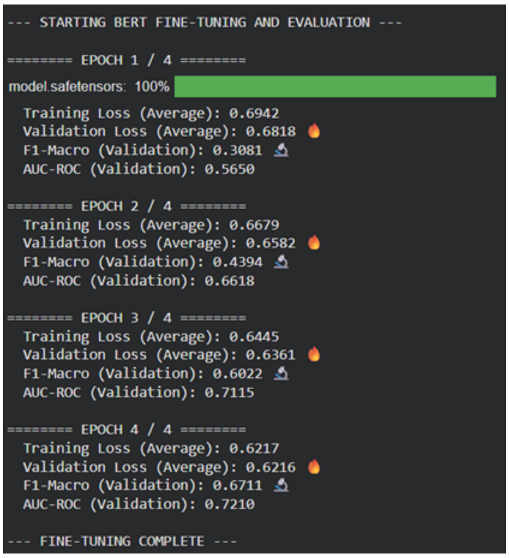
Classification and fine-tuning In Fig. 6. is shown the result of fine-tuning by the epoch hyperparameter and using the metrics average F1-Score, AUC_ROC and it was checked the loss. The F1-Macro score is used when there are many labels and is an average of the 36 individual F1-scores. The AUC-ROC is used to measure the efficient separation between classes. The training loss is used to determine the discrepancy between the model’s predictions and the actual values, while the validation loss is used to determine its ability to generalize. If the model begins to memorize the training data too much (overfitting), the training loss will continue to decrease, but the validation loss will start to increase or plateau. As can be noted, the Training Loss decreases from 0.6942 (Epoch 1) to 0.6217 (Epoch 4). The Validation Loss decreases from 0.6818 to 0.6216, and in Epoch 4, the Training Loss (0.6217) and the Validation Loss (0.6216) are nearly identical, suggesting that the model has learned efficiently from the training data without losing its ability to generalize to the validation data.The F1-Macro (Validation) at 0.6711 (67.11%) evaluates the model’s accuracy in identifying each of the 36 symptoms individually. This value is not the best achievable, but only 1000 samples were used for training, and this affected finding a better fit. The collection time for 1000 samples is three and a half hours.The AUC-ROC (Validation) at 0.7210 (72.10%) indicates good separation between the classes. A value of 0.72 is a solid performance, especially for a complex clinical task with a low quantity of training data, as previously mentioned. The process was not extended to more data due to the runtime implications of even a single change to the hyperparameters. In fact, the 1000 samples were only saved once the model had stabilized for learning (for example, no longer crashing). Minor adjustments to hyperparameters and configurations were made with 100 samples, and then the dataset with all 1000 samples was saved. Since only one computer was used, runtime disconnections and server crashes prevented expanding the dataset and improving the model fit.

Differential Diagnosis Function (Cosine Similarity)

A reference vector was created from Table 1. These are the symptoms summarized in 5 areas of impact. These vectors serve as focused pure vectors, as explained previously. The vector is calculated using the trained model; the same process is performed for the user vector, which is described below:

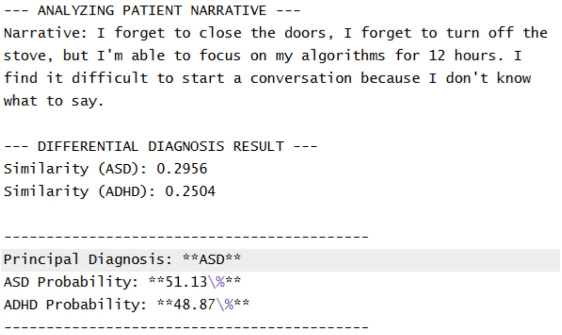
patient_narrative = "I forget to close the doors, I forget to turn off the stove, but I'm able to focus on my algorithms for 12 hours. I find it difficult to start a conversation because I don't know what to say."

The preceding fragment would describe an autistic/Asperger profile in a very basic way. The results of the similarity cosine between patient_narrative and the reference vector for both disorders are shown in Fig. 7.



```
--- STARTING BERT FINE-TUNING AND EVALUATION ---

======== EPOCH 1 / 4 ========
model.safetensors: 100%
  Training Loss (Average): 0.6942
  Validation Loss (Average): 0.6818 🔥
  F1-Macro (Validation): 0.3081 🔬
  AUC-ROC (Validation): 0.5650

======== EPOCH 2 / 4 ========
  Training Loss (Average): 0.6679
  Validation Loss (Average): 0.6582 🔥
  F1-Macro (Validation): 0.4394 🔬
  AUC-ROC (Validation): 0.6618

======== EPOCH 3 / 4 ========
  Training Loss (Average): 0.6445
  Validation Loss (Average): 0.6361 🔥
  F1-Macro (Validation): 0.6022 🔬
  AUC-ROC (Validation): 0.7115

======== EPOCH 4 / 4 ========
  Training Loss (Average): 0.6217
  Validation Loss (Average): 0.6216 🔥
  F1-Macro (Validation): 0.6711 🔬
  AUC-ROC (Validation): 0.7210

--- FINE-TUNING COMPLETE ---
```

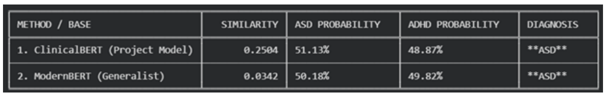
Fig. 6.



```
--- ANALYZING PATIENT NARRATIVE ---
Narrative: I forget to close the doors, I forget to turn off the
stove, but I'm able to focus on my algorithms for 12 hours. I
find it difficult to start a conversation because I don't know
what to say.

--- DIFFERENTIAL DIAGNOSIS RESULT ---
Similarity (ASD): 0.2956
Similarity (ADHD): 0.2504

-------------------------------------------
Principal Diagnosis: **ASD**
ASD Probability: **51.13\%**
ADHD Probability: **48.87\%**
-------------------------------------------
```

Fig. 7.

Code was added to prepare the loading and execution of another model, ModernBERT. This model was suggested by the evaluation team for comparative purposes. The result was as follows:



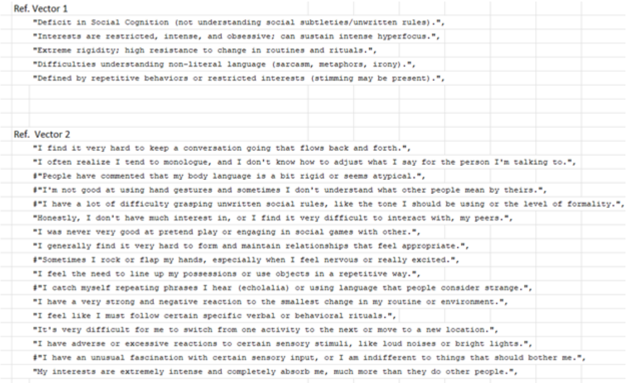
```
| METHOD / BASE                  | SIMILARITY | ASD PROBABILITY | ADHD PROBABILITY | DIAGNOSIS |
| 1. ClinicalBERT (Project Model)|     0.2504 | 51.13%          | 48.87%           | **ASD**   |
| 2. ModernBERT (Generalist)     |     0.0342 | 50.18%          | 49.82%           | **ASD**   |
```

Fig. 8.

The value of 0.2504 is significantly higher than the generalist model. This suggests that ClinicalBERT's predictions are four times more similar to the correct answer. ModernBERT's similarity value was 0.034. This demonstrates that the fine-tuning of ClinicalBERT has allowed it to generate text representations that are much more relevant and specific for the clinical task. The 0.034 value is low for the specialized task. In this test, both models agree on the diagnostic result, suggesting that the narrative likely corresponds to a person with ASD. This is true, given the symptoms described in the narrative. This agreement in results can change drastically if the reference vector of symptoms is modified, even when using the same user narrative. This will be demonstrated in the following section.

Analysis using different types of reference values

Fig. 9. shows two types of reference vectors. The first corresponds to the symptoms in Table 1. For convenience, only the symptoms of ASD (Autism Spectrum Disorder) are shown. The complete code contains the criteria for ADHD (Attention Deficit Hyperactivity Disorder) and can be found in the repository. Reference Vector 1 (Ref. Vector 1) condenses the reference vector representation in a more compact form. Reference Vector 2 (Ref. Vector 2) contains the 18 ASD symptoms in colloquial language, reflecting practical situations. This second vector will make the vector representation space more diffuse. There is a third reference vector in the code that is not shown here but can be executed using the code from the repository. The conclusion to be highlighted is that the mathematical representation (embedding) is greatly affected by the narrative of the reference vectors; it can even completely alter the diagnostic outcome, as shown in Fig. 10.

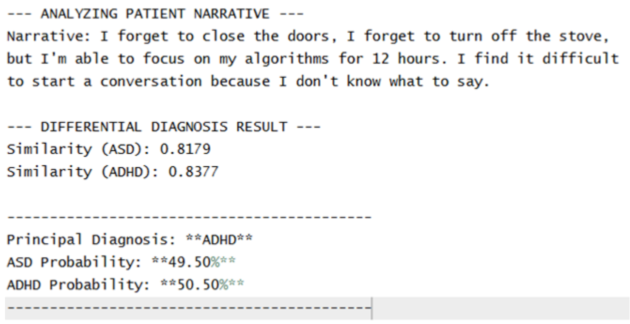


```
Ref. Vector 1
    "Deficit in Social Cognition (not understanding social subtleties/unwritten rules).",
    "Interests are restricted, intense, and obsessive; can sustain intense hyperfocus.",
    "Extreme rigidity; high resistance to change in routines and rituals.",
    "Difficulties understanding non-literal language (sarcasm, metaphors, irony).",
    "Defined by repetitive behaviors or restricted interests (stimming may be present).",



Ref. Vector 2
    "I find it very hard to keep a conversation going that flows back and forth.",
    "I often realize I tend to monologue, and I don't know how to adjust what I say for the person I'm talking to.",
    #"People have commented that my body language is a bit rigid or seems atypical.",
    #"I'm not good at using hand gestures and sometimes I don't understand what other people mean by theirs.",
    #"I have a lot of difficulty grasping unwritten social rules, like the tone I should be using or the level of formality.",
    "Honestly, I don't have much interest in, or I find it very difficult to interact with, my peers.",
    "I was never very good at pretend play or engaging in social games with other.",
    "I generally find it very hard to form and maintain relationships that feel appropriate.",
    #"Sometimes I rock or flap my hands, especially when I feel nervous or really excited.",
    "I feel the need to line up my possessions or use objects in a repetitive way.",
    #"I catch myself repeating phrases I hear (echolalia) or using language that people consider strange.",
    "I have a very strong and negative reaction to the smallest change in my routine or environment.",
    "I feel like I must follow certain specific verbal or behavioral rituals.",
    "It's very difficult for me to switch from one activity to the next or move to a new location.",
    "I have adverse or excessive reactions to certain sensory stimuli, like loud noises or bright lights.",
    #"I have an unusual fascination with certain sensory input, or I am indifferent to things that should bother me.",
    "My interests are extremely intense and completely absorb me, much more than they do other people.",
```

Fig. 9.

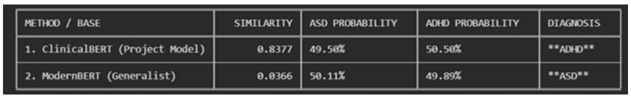
https://github.com/mhernandezlordui/Final-Project-LLM

For the test, some symptoms were removed from the narrative of Reference Vector 2 (for both disorders), while the user's narrative remained the same, which, as mentioned, corresponds to an individual with ASD. This change in the reference vector representation modified the similarity calculation values and, consequently, the diagnosis. Note that the results in Fig. 10 show a different diagnosis. ClinicalBERT failed, and ModerBERT was correct (See Fig.11). This result is very interesting because it demonstrates that the diagnosis cannot be made using the cosine of similarity due to its high sensitivity. However, it introduces the intuitive idea that a deep reasoning stage of LLM or a Reinforcement Learning Model should be incorporated. Therefore, ClinicalBERT embeddings are good for capturing the semantics of the clinical narrative, but the Cosine Similarity method is insufficient to handle the complexity and hierarchy of a real differential diagnosis.



```
--- ANALYZING PATIENT NARRATIVE ---
Narrative: I forget to close the doors, I forget to turn off the stove,
but I'm able to focus on my algorithms for 12 hours. I find it difficult
to start a conversation because I don't know what to say.

--- DIFFERENTIAL DIAGNOSIS RESULT ---
Similarity (ASD): 0.8179
Similarity (ADHD): 0.8377

-------------------------------------------
Principal Diagnosis: **ADHD**
ASD Probability: **49.50%**
ADHD Probability: **50.50%**
-------------------------------------------
```

Fig. 10.



METHOD / BASE	SIMILARITY	ASD PROBABILITY	ADHD PROBABILITY	DIAGNOSIS
1. ClinicalBERT (Project Model)	0.8377	49.50%	50.50%	**ADHD**
2. ModernBERT (Generalist)	0.0366	50.11%	49.89%	**ASD**

Fig. 11.

8 Conclusion

This project successfully demonstrated the feasibility of using a clinical Large Language Model (LLM) for complex classification and vector representation tasks in the domain of psychiatric diagnosis (ASD and ADHD). However, it also identified a critical limitation in purely mathematical methodologies. The advanced capabilities of the Gemini model were used in a Few-Shot Setting to generate a synthetic dataset of 1000 clinical narratives, each labeled with 36 symptoms (ASD/ADHD). Diagnosis by pure vector comparison is insufficient and limited by high sensitivity. Tests showed that Cosine Similarity is too sensitive to the construction of the reference vector (whether it was compact versus narrative) and to small variations in the patient's input narrative (e.g., removing a symptom). This resulted in a direct and potentially incorrect modification of the final diagnosis. ClinicalBERT embeddings are excellent for capturing the deep semantics of the clinical narrative, but the decision stage requires a more sophisticated process.

9 AI Disclosure

- Did you use any AI assistance to complete this proposal? If so, please also specify what AI you used.
  - I used Gemini assistance as described in some parts of this document.

If you answered yes to the above question, please complete the following as well:

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.
  - I have text created with the DIVA, AQ-10, EQ, AAA, etc. tests and I have 36 labels obtained from the DSM-5 criteria in a CSV file. How could I augment the data for BERT?
  - How can I configure the Gemini Client API for text generation?
  - How many samples should I have in my augmented dataset for the model to train properly?
  - Using the labels from photo 1 and the ASD and ADHD criteria from the DSM-5, create a description like this example: ASD_Reciprocity: Deficits in socio-emotional reciprocity (difficulty initiating or maintaining a conversation, lack of interest in peers).
  - Code for setting clinicalBERT as a multi-label classifier.
  - All conversations were conducted in my native language, Spanish.
- **Free response:** In my opinion, AI assistance is fantastic. It saves a lot of time in the exploration and library setup phase. If you

https://github.com/mhernandezlordui/Final-Project-LLM

have a clear understanding of your project's functionality, AI can assist with code examples that you can then adapt to your specific project.

References

Autism spectrum disorder in adults: diagnosis and management (cg142).

2012. Autism spectrum disorder in adults: diagnosis and management (cg142). Consultado en la sección Recomendaciones.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition. American Psychiatric Publishing, Arlington, VA. Consultado en línea.

Kevin M Antshel and Natalie Russo. 2019. Autism spectrum disorders and adhd: Overlapping phenomenology, diagnostic issues, and treatment considerations. *Current psychiatry reports*, 21(5):34.

Simon Baron-Cohen, Sally Wheelwright, Richard Skinner, John Martin, and Rachael Clubley. 2006. The adult asperger assessment (aaa): A diagnostic method. *Journal of Autism and Developmental Disorders*, 36(7):871–880.

Chien Wen Chien and Yueh-Ming Tai. 2024. Performances of large language models in detecting psychiatric diagnoses from chinese electronic medical records: Comparisons between gpt-3.5, gpt-4, and gpt-4o. *Taiwanese Journal of Psychiatry*, 38(3):134–141.

Laura Fusar-Poli, Natascia Brondino, Pierluigi Politi, and Eugenio Aguglia. 2022. Missed diagnoses and misdiagnoses of adults with autism spectrum disorder. *European archives of psychiatry and clinical neuroscience*, 272(2):187–198.

Christopher Gillberg, Carina Gillberg, Maria Råstam, and Elisabeth Wentz. 2001. The asperger syndrome (and high-functioning autism) diagnostic interview (asdi): A preliminary study of a new structured clinical interview. *Autism*, 5(1):57–66.

J. J. S. Kooij and M. H. Francken. 2019. *Diagnostic Interview for ADHD in Adults (DIVA 5)*. DIVA Foundation. Versión en español consultada en línea.

Johnny Matson, Robert Rieske, and Lindsey Williams. 2013. The relationship between autism spectrum disorders and attention-deficit/hyperactivity disorder: An overview. *Research in developmental disabilities*, 34:2475–2484.

Tamara May, Pamela D Pilkington, Rita Younan, and Katrina Williams. 2021. Overlap of autism spectrum disorder and borderline personality disorder: a systematic review and meta-analysis. *Autism Research*, 14(12):2688–2710.

Susan Dickerson Mayes, Susan L Calhoun, Rebecca D Mayes, and Sarah Molitoris. 2012. Autism and adhd: Overlapping and discriminating symptoms. *Research in Autism Spectrum Disorders*, 6(1):277–285.

Prasenjit Mukherjee, RS Gokul, Sourav Sadhukhan, Manish Godse, and Baisakhi Chakraborty. 2023. Detection of autism spectrum disorder (asd) from natural language text using bert and chatgpt models. *International Journal of Advanced Computer Science and Applications*, 14(10).

Craig J Newschaffer, Lisa A Croen, Julie Daniels, Ellen Giarelli, Judith K Grether, Susan E Levy, David S Mandell, Lisa A Miller, Jennifer Pinto-Martin, Judy Reaven, and 1 others. 2007. The epidemiology of autism spectrum disorders. *Annual review of public health*, 28(1):235–258.

Xiaoyu Ren, Yuanchen Bai, Huiyu Duan, Lei Fan, Erkang Fei, Geer Wu, Pradeep Ray, Menghan Hu, Chenyuan Yan, and Guangtao Zhai. 2023. Chatasd: Llm-based ai therapist for asd. In *International Forum on Digital TV and Wireless Multimedia Communications*, pages 312–324. Springer.

Riva A. Ritvo, Edward R. Ritvo, Donald Guthrie, Eric R. Ritvo, Vicki Hufnagel, William McMahon, Paula Toniolo, and Daniel Morris. 2011. The ritvo autism asperger diagnostic scale-revised (raads-r): A scale to assist the diagnosis of autism spectrum disorder in adults. *Journal of Autism and Developmental Disorders*, 41(8):1119–1136.

Emily Simonoff, Andrew Pickles, Tony Charman, Susie Chandler, Tom Loucas, and Gillian Baird. 2008. Psychiatric disorders in children with autism spectrum disorders: prevalence, comorbidity, and associated factors in a population-derived sample. *Journal of the American Academy of Child & Adolescent Psychiatry*, 47(8):921–929.

Regina Taurines, Christina Schwenck, Eva Westerwald, Michael Sachse, Michael Siniatchkin, and Christine Freitag. 2012. Adhd and autism: differential diagnosis or overlapping traits? a selective review. *ADHD Attention Deficit and Hyperactivity Disorders*, 4(3):115–139.