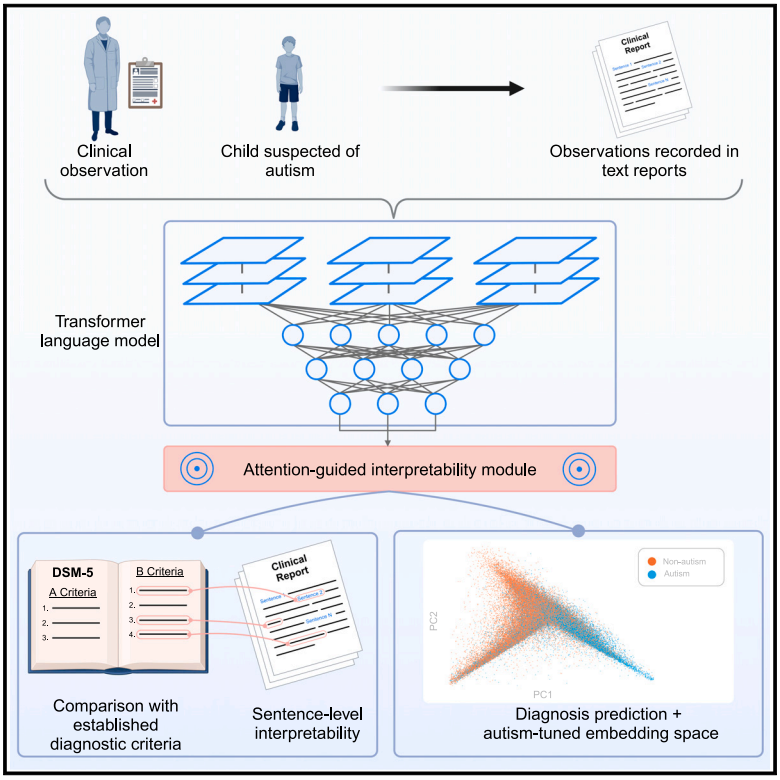
Cell

Large language models deconstruct the clinical intuition behind diagnosing autism

Graphical abstract



Highlights

- Language models enable a robust autism diagnosis from healthcare text reports alone
- Interpretability framework points to autism-relevant sentence elements from reports
- Comparison against DSM-5 criteria for each report, although not measured in subjects
- Repetitive behaviors emerged as more salient for autism diagnosis than social factors

Authors

Jack Stanley, Emmett Rabot, Siva Reddy, Eugene Belilovsky, Laurent Mottron, Danilo Bzdok

Correspondence

danilo.bzdok@mcgill.ca

In brief

Clinical assessment is the gold standard for autism diagnosis, and by using amply pre-trained language models paired with a suitable interpretability strategy, the native analysis of clinical intuition itself is now possible, guided by detailed observational text reports from clinicians. Our interpretable language model framework identified repetitive and stereotyped behaviors as being empirically more relevant for autism diagnosis than social signs in the reports, motivating the possible revision of existing diagnostic instruments.

Stanley et al., 2025, Cell *188*, 2235–2248
April 17, 2025
https://doi.org/10.1016/j.cell.2025.02.025

CellPress

Cell

CellPress
OPEN ACCESS

Article

Large language models deconstruct the clinical intuition behind diagnosing autism

Jack Stanley,[1,2,6] Emmett Rabot,[3,4,6] Siva Reddy,[1] Eugene Belilovsky,[1,5] Laurent Mottron,[3,4,7] and Danilo Bzdok[1,2,7,8,*]
[1]Mila - Québec Artificial Intelligence Institute, Montréal, QC H2S3H1, Canada
[2]The Neuro - Montréal Neurological Institute (MNI), McConnell Brain Imaging Centre, Department of Biomedical Engineering, Faculty of Medicine, School of Computer Science, McGill University, Montréal, QC H3A2B4, Canada
[3]Research Center, Centre Intégré Universitaire de Santé et de Services Sociaux du Nord-de-l'Ile-de-Montréal (CIUSSS-NIM), Montréal, QC H4K1B3, Canada
[4]Université de Montréal, Montréal, QC H3C3J7, Canada
[5]Department of Computer Science and Software Engineering, Concordia University, Montreal, QC H3G 1M8, Canada
[6]These authors contributed equally
[7]These authors contributed equally
[8]Lead contact
*Correspondence: danilo.bzdok@mcgill.ca
https://doi.org/10.1016/j.cell.2025.02.025

SUMMARY

Efforts to use genome-wide assays or brain scans to diagnose autism have seen diminishing returns. Yet the clinical intuition of healthcare professionals, based on longstanding first-hand experience, remains the gold standard for diagnosis of autism. We leveraged deep learning to deconstruct and interrogate the logic of expert clinician intuition from clinical reports to inform our understanding of autism. After pre-training on hundreds of millions of general sentences, we finessed large language models (LLMs) on >4,000 free-form health records from healthcare professionals to distinguish confirmed versus suspected autism cases. By introducing an explainability strategy, our extended language model architecture could pin down the most salient single sentences in what drives clinical thinking toward correct diagnoses. Our framework flagged the most autism-critical DSM-5 criteria to be stereotyped repetitive behaviors, special interests, and perception-based behaviors, which challenges today's focus on deficits in social interplay, suggesting necessary revision of long-trusted diagnostic criteria in gold-standard instruments.

INTRODUCTION

1%–2% of individuals in our societies probably meet our current diagnostic criteria for autism.[1] In the quest toward objective diagnostic indicators for autism, researchers have ventured to identify biomarkers in complex biological data, which are epitomized by common-variant genetics profiles (i.e., genome-wide association study [GWAS]) and brain-imaging recordings (i.e., magnetic resonance imaging [MRI])—two of the most important and widely used high-throughput technologies that have emerged in biomedicine.[2–4] Despite laborious and expensive attempts, the use of biomarkers to diagnose autism has not materialized in everyday clinical reality so far. Our failure to identify reliable objective autism markers speaks to our limited understanding of autism itself—an alternative research paradigm may be in order.

In particular, there has been a focus on establishing reliable genetic markers for autism. Despite these efforts, the diagnostic yield, which denotes the rate of true positives identified by a diagnostic tool, of genetic tests for autism has remained well below the threshold needed for clinical adoption.[5] Autism is known to be highly heritable. Yet, as evidenced by twin and family studies,[6] the genetic architecture of the disorder is also highly polygenic: conceivably tens of thousands of associated single-nucleotide polymorphisms (SNPs)[7] are thought to contribute to the phenotypic presentation in concert. To address this apparent polygenicity, polygenic risk scores (PRSs) have become an increasingly common tool for the assessment of an individual's genetic propensity for certain brain disorders, including autism.[8] PRS involves computing a weighted sum of potentially tens of thousands of autism-associated genetic variants, as determined through GWAS.[4] A large-scale study evaluated the predictive ability of their autism PRS model as accounting for a modest ~2% fraction of the observed variance of the autism trait.[9]

The exploration of predictive machine learning algorithms for diagnosing autism from combinations of individual genetic loci has gained momentum in the 2000s. One such investigation included 237 SNPs located in 146 target genes as input fed into their predictive model for autism.[10] Their hope was that the machine learning approach would fully exploit subtle genetic differences to automatically extract diagnosis-relevant patterns in combinations of individual SNPs. These authors reported a

diagnostic classification accuracy of 71.6% on balanced yet ethnically homogeneous independent test datasets. However, this accuracy turned out to be even weaker at only 56.4% when the same predictive model was evaluated on individuals that were ethnically dissimilar to those in the training set. This vexing state of affairs is not entirely unexpected, and there has been mild progress in identifying "autism-specific" genes[11] or dedicated gene pathways[12] that would point to mechanisms that are idiosyncratic to the primary biology of autism. Indeed, the foundation of heritable autism traits is distributed across numerous marginal genetic distinctions in the genome.[13] For these reasons, the use of today's genetic testing procedures for individuals with suspected autism is unlikely to yield clinically useful insight soon.

Even earlier, since the 1990s, another pervasive technique for the automatic diagnosis of autism is the use of brain-imaging scanners that can non-invasively record from the central nervous system. For example, in a seminal benchmarking study, Abraham and colleagues employed resting-state functional MRI (R-fMRI) to classify individuals from the largest-of-its-kind dataset (Autism Brain Imaging Data Exchange [ABIDE]) into autism or neurotypical control groups.[14] In a rigorous, comprehensive machine learning assessment, these authors aimed to estimate the best possible prediction performance for the diagnosis of autism from brain-imaging data. Their careful benchmark concluded that the support vector classifier with L2 penalization yielded the highest autism prediction performance on independent test individuals, at 66.8% classification accuracy—an improvement over previous attempts at R-fMRI-based diagnosis classification from the ABIDE dataset at 60% accuracy.[15] A later study that used deep learning to improve upon previous diagnostic classification saw a modest performance increase, with an accuracy of 70% on the ABIDE cohort.[16] The documented lack of success in clinical translation from brain imaging in autism may be due to a combination of reasons.[17,18] As autism studies have scaled up to deploy ever more powerful predictive modeling techniques, it is becoming clear that there may be limited information about autism that can be gleaned from brain scans alone.

The investigation of data modes with high intrinsic information density as to autism, motivating our present approach, will be unavoidable for deepening our diagnostic comprehension of the disorder. With today's paucity of viable autism biomarkers, be they genetic profiles, brain-imaging measurements, blood samples, or any other body-derived metric that we have explored, the burden of diagnosis rests largely on healthcare practitioners. Clinical judgment involves providing qualitative evaluation of individuals based on the alignment of observations to established diagnostic criteria. Autism presents with a wide variety of symptoms and severity levels. Therefore, healthcare practitioners today have no other choice in reserve than trusting their clinical intuition, honed through years of training and first-hand experience, to reach an accurate diagnosis. This accumulated clinical know-how should not be overlooked and should instead be considered an indispensable repository of knowledge from which we can draw implementable conclusions. Furthermore, the clinical thought process, by definition, represents the most compact and high-fidelity source of diagnosis-relevant information pertaining to autism. However, unmoored from their initial foundations, the implementation of the diagnostic guidelines has drifted significantly, with only half the symptom items required for diagnosis today compared with 20 years ago.[19]

There is hence a lack of actionable insight from biomarkers yet undeniable practical success driven by the clinical intuition of healthcare practitioners. Therefore, we here tried an unconventional approach: we extracted and interrogated this clinical intuition to more directly inform our understanding of autism and how to diagnose it. To accomplish this goal, we capitalized on a cohort of >1,000 children from a large, demographically representative population spanning a broad geographic area who were all referred because of suspected autism, totaling to >4,000 digital health records from clinical practitioners. We treated these detailed records as demonstrations of the clinical thought process itself—assuming that human thought is reflected in human language[20]—representing an untapped treasure trove of knowledge. Fortunately, recent advances in natural language processing (NLP), culminating in the development of pre-trained large language models (LLMs), have pushed forward the direct quantification and analysis of unstructured text[20]—abundant in our healthcare systems, yet currently underexploited.

By developing a mission-tailored language model architecture, we interrogated the individual semantic elements that exert profound influence on the clinical process of autism diagnosis. This language model, pre-trained on hundreds of millions of general language sentences, was refined on our corpus of clinical text reports to achieve a robust diagnostic classification accuracy on independent reports, demonstrating that diagnosis-relevant signal can be successfully extracted by transformer language models. More importantly, our purpose-built language model architecture was designed to be natively interpretable. We use this framework to unpack the most essential drivers of autism diagnosis as articulated by the experienced clinicians themselves. With the derived context of this autism-aware space, we assessed the relative usefulness of each specific Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5)[21] autism criterion in contributing to the establishment of a precise diagnosis. We thus showed that the clinical thought process itself can be deconvolved to provide actionable insight into what the diagnostic criteria of autism should be.

RESULTS

Rationale and analysis workflow

In contrast to consensus research in autism, our study began from the assertion that there is untapped value in dissecting the clinical expertise of healthcare professionals frequently exposed to large, diverse populations enriched in autism. We expected that clinical thinking could be deconvolved through a fully data-driven analysis of text reports to lend insights into the essence of what distinguishes observations of a child with autism from one in whom a diagnosis has been clinically ruled out. In a first step, we tested whether leveraging the advanced text processing capabilities of large pre-trained language models could extract valuable information for precisely predicting autism diagnosis, based solely on clinical observations of

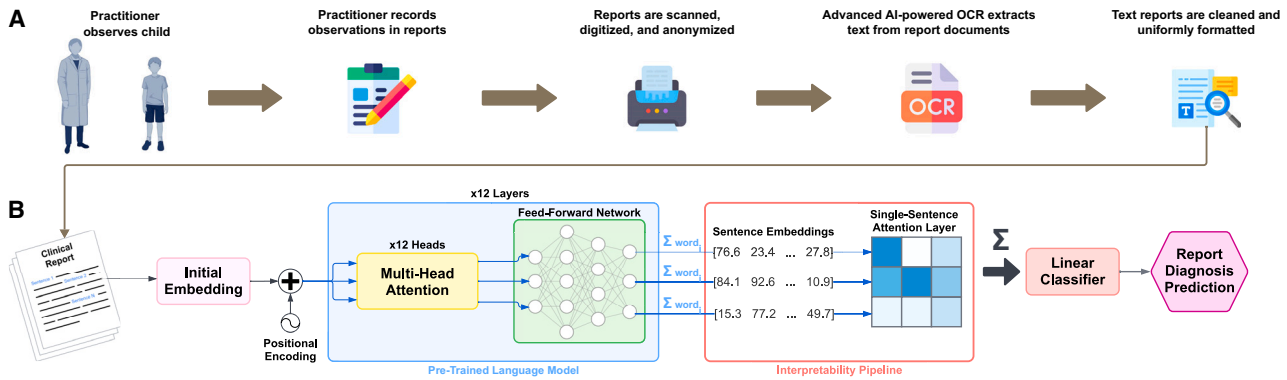


Figure 1. Language model technology can locate sentences critical for autism diagnosis in digitized medical records

(A) >4,000 hard-copy free-form healthcare professional reports, in French, of >1,000 children assessed over a 4-year period were scanned and converted to PDF format, and text was parsed: we used a computer vision optical character recognition (OCR) pipeline to convert each health record into a plain text format that was fully actionable on language model frameworks.

(B) We facilitated sentence-level analysis by feeding each sentence from each report individually into a French-language RoBERTa-based language model backbone pre-trained on 12.8 billion word tokens of general-purpose text. This encoded the sentences into a semantically rich and quantitatively comparable embedding space. Individual words are tokenized in a format that is actionable by the model, subsequently embedded as dense vectors in a continuous space by our pre-trained model, and combined with positional encodings that capture the order of the words in the sequence. Sentence embeddings are constructed by mean-pooling the word embedding representations produced by our pre-trained language model, forming the fundamental unit of our interpretability pipeline. The next processing component consists of a trainable single-head attention module that is (jointly) trained to pinpoint and preferentially weight the language model-embedded sentences that are most helpful in diagnosis classification, thereby automatically identifying salient sentence-level information from each healthcare professional report. The overall report embedding is then constructed as a weighted average of the sentence embeddings. The final module is a linear classification layer that predicts the final consensus diagnosis based on the attention-guided report embedding. The two-component model, including the pre-trained backbone and the single-head attention layer, was fine-tuned on the >4,000 reports in an end-to-end fashion to predict the diagnosis. This fine-tuning process shapes the embedding space into an autism-aware semantic space, where each sentence embedding reflects its relationship to a potential autism diagnosis.

individuals, without checking DSM-5 criteria. Next, we sought to rigorously understand and quantify the most relevant elements in reports from healthcare professionals during the diagnostic process that led up to confirmed autism diagnoses, as conveyed by healthcare professionals experienced in working with children with autism. Further, we realized that centering the single-sentence level as the unit of investigation (as opposed to the word level or report level) would enable a high degree of granularity for downstream interpretation while retaining outstanding predictive accuracy of the diagnosis. To enable sentence-level granularity in analysis and interpretation, we architected a trainable single-head attention module as the final layer of our language model framework. This single-head attention module pinpointed and selectively flagged the most autism-critical sentences in each report. In short, the pre-trained language model leverages transfer learning to extract relevant semantic elements from the text, while the single-head attention module enables direct interpretability of these semantic elements.

Language models extract underlying semantic nuances of autism features from clinical health records

We devised a pipeline using a pre-trained general-purpose language model, followed by fine-tuning on our domain-specific corpus. Specifically, we deployed a RoBERTa-based[22] language model (138 million model parameters in total) that was pre-trained on 12.8 billion word tokens, totaling to approximately 489 million training sentences of broad-domain text corpora. This pre-training regimen spawned a general-purpose language representation in the model,[20] which enabled us to effectively use transfer learning to fine-tune the model toward predicting autism diagnosis on our carefully collected clinical dataset of 4,272 complete autism-focused reports from 1,080 different patients, each with potentially several clinical visits (Figure 1).

The entire model was fine-tuned end-to-end on our collection of 4,272 reports with the goal of classifying the associated autism diagnosis for each report: clinically confirmed versus suspected but ruled-out autism case. This end-to-end fine-tuning entailed further training the parameters of the pre-trained language model, the single-head attention module, and the final classification layer jointly on the autism diagnosis classification objective. We calculated the classification performance accuracies following a rigorous 5-fold cross-validation framework: for each cross-validation fold, 80% of the reports were randomly selected to compose the training set on which our model was fine-tuned, while the remaining 20% of new unseen reports constituted the test set on which the out-of-sample accuracies for each fold were calculated. To respect independence between individuals, multiple reports from the same patient were exclusively allocated to either the test set or the training set in each fold.

In comparison to standard bag-of-words (BOW) approaches and a more advanced Doc2Vec algorithm[23] for diagnosis classification, which do not benefit from pre-training, our language model framework robustly outperformed in terms of average out-of-sample classification accuracy (Figure 2A). Our language model achieved an average accuracy of 79.4% (SD, 0.9%) on

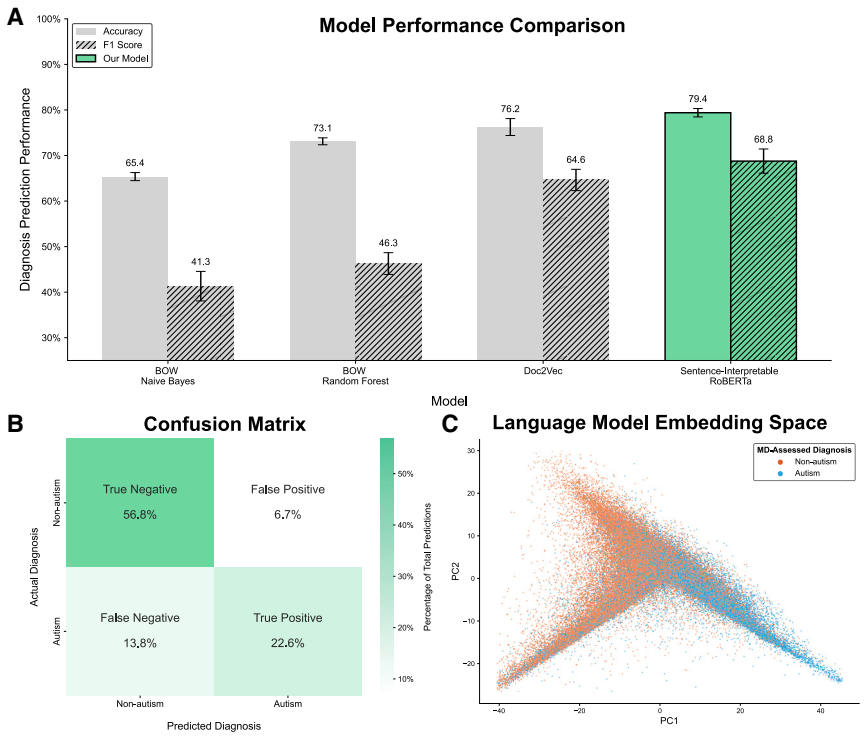


Figure 2. Pre-trained language models enable robust classification accuracy of autism diagnoses from healthcare professional reports

(A) Out-of-sample autism diagnosis prediction accuracy from unstructured health professional reports, from traditional and more advanced natural language processing algorithms. Bar height indicates the average score after 5-fold cross-validation, and whiskers indicate the variability (one standard deviation) of model prediction performance after fitting each model through 5-fold cross-validation. Solid-colored bars correspond to raw classification accuracy, while bars with diagonal stripes correspond to F1 score. Our model is shaded green. All models were trained on the same corpus of reports and evaluated using the same 5-fold cross-validation scheme. Our single-sentence pre-trained RoBERTa-based model, fine-tuned on our corpus of text reports, shows higher average classification accuracy compared with traditional NLP approaches, in addition to allowing for greater interpretability. See Figure S1 and Table S1 for additional benchmarks. BOW, bag of words.

(B) Confusion matrix of our single-sentence RoBERTa-based model predictions. The vertical axis corresponds to the actual clinically designated autism diagnosis, while the horizontal axis denotes our model predictions for the diagnosis label, per report.

(C) Our fine-tuned language model spans an autism-aware embedding space at single-sentence granularity. PCA decomposition of sentence-level embeddings generated by our language model reveals a separation of sentences taken from clinically confirmed autism cases versus non-autism cases. Each point in the PCA plot corresponds to a single sentence and is colored according to the final diagnosis label: autism refers to clinically confirmed autism cases, and "non-autism" refers to suspected but clinically ruled-out cases. The learned underlying embedding space instantiated by our transformer language model allows for direct comparison and contrast between any two natural language sentences, regardless of origin, regarding information value for autism diagnosis.

See also Figure S1.

unseen reports, while the baseline BOW models achieved an accuracy of only 65.4% (SD, 0.9%) with a linear naive Bayes classifier and 73.1% (SD, 0.8%) with a non-linear random forest classifier. The Doc2Vec method yielded a classification accuracy of only 76.2% (SD, 1.9%). Therefore, our fine-tuned language model, enhanced by transfer learning, enabled superior generalization by a higher diagnosis classification accuracy compared with traditional NLP methods (see Figure 2B for a detailed confusion matrix of our model predictions). What is essential for interpretability is our model's ability to distinguish between clinically confirmed and suspected but ruled-out cases of autism, rather than the raw autism detection rate. Similar classification performance was observed with other transformer-based models, including Longformer,[24,25] Llama 3.1 8B,[26] and Gemma 7B[27] (see Figure S1 and Table S1 for extended benchmarks). Consequently, our fine-tuned language model was able to effectively extract relevant information from unstructured clinical notes and thus also aspects of the clinical thought process that allowed for reliable autism diagnosis detection.

Unpacking language model internals at sentence-level granularity

After ascertaining our model's ability to correctly predict autism diagnoses, underscoring the disorder-relevance of its semantic representation space, we leveraged our sentence-level strategy to open a window into the internals of what is commonly considered to be an impenetrable black box.[20] Our objective was to conduct an initial confirmatory assessment that the internal sentence-level representation of the language model indeed captured aspects related to individual clinical features that characterize autism. We extracted the major factors of variation (using principal-component analysis [PCA]) across all hidden dimensions of the sentence embeddings for each sentence in our corpus into a more compact 2D representation (Figure 2C). Hence, we effectively brought to the surface important aspects of the internal semantic representational space spanned by our model. Each data point in the PCA analysis (dot in scatterplot, Figure 2C) corresponds to an individual sentence, colored post-hoc according to the diagnosis associated with the report from which each sentence originates. As evidenced by this latent space exploration, after we fine-tuned the model on our collection of autism reports, the language model spanned a meaningful autism-aware embedding space: sentences carrying information associated with autism diagnosis are separated from sentences that convey semantic elements that were unhelpful for the diagnosis of autism. Moreover, this autism-sensitive semantic representation covers the entirety of possible natural language inputs, meaning that it allows for the comparison of any two sentences, regardless of their origin, based on their relevance for autism classification (cf. below).

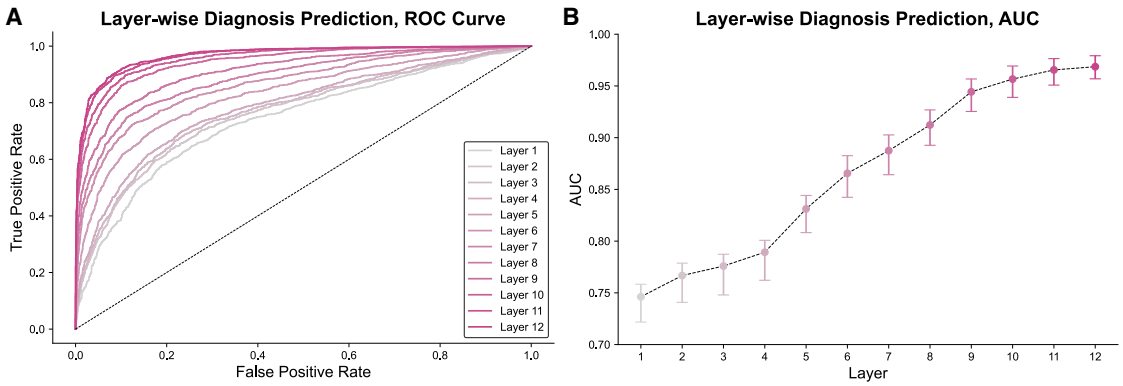


Figure 3. Layer-wise prediction of autism diagnosis from report embeddings speaks to how language model makes decisions internally

(A) Receiver operating characteristic (ROC) curve for the diagnosis prediction task from the pooled report embeddings for each layer of our fine-tuned model. As information is flowing through our model architecture's processing layers, the pooled report embeddings become increasingly useful in predicting the diagnosis. As the model extracts and combines more relevant information from every part of the report at each layer, the pooled embeddings become increasingly useful for distinguishing diagnosed autistic from non-autistic subjects. The depth of the chosen pre-trained model appears critical for the accurate classification of the diagnosis. It is possible to extract internal language model features that are increasingly relevant to the classification task. Each curve for each layer was averaged over 5 cross-validation folds.

(B) Analogous to (A), but shows the prediction performance (average area under the curve [AUC]) for the autism prediction models layer by layer, with error bars indicating the standard deviation across 5 cross-validation folds.

To dissect the viscera of this multi-layer deep learning architecture in yet another complementary way, we aimed to trace properties of the internal logic of the language model as the text report information is transformed and moved through each layer of the model. After mean-pooling the individual sentence embeddings for each report at each of the 12 layers of the deep neural network, we used this report-level embedding as a basis to train a predictive linear classifier (logistic regression) on the associated diagnosis label. We found that there was a steady increase in the autism classification performance as information from a report moved deeper and deeper into the model (Figures 3A and 3B): starting in the 1st layer with an average area under the curve (AUC) of 0.746 (SD, 0.014), in the middle 6th layer with an average AUC of 0.866 (SD, 0.014), and finally reaching its maximum performance at the final 12th layer with an AUC of 0.968 (SD, 0.008) averaged across cross-validation folds. Thus, these analyses confirmed that the sentence embeddings successively transformed by the model layers become increasingly diagnosis-relevant as the model extracts and recombines the most essential information from each report for the diagnosis classification task. Indeed, it is this highly informative distillation from layer 12 that was fed directly into our single-sentence attention module, where the most important sentences for the diagnosis are automatically identified and preferentially up-weighted. Hence, our devised single-sentence attention module served to refine the signal that the pre-trained language model reads from the entire report content for the sake of human comprehension.

The use of a single-head attention module as a trainable filter inside our language model allows for a naturally interpretable device for weighing each sentence's importance within a report in relation to the diagnosis prediction. The sentences that turned out to be most strongly attended were assigned greater weight in the pooled report embedding. Therefore, these sentences, which prove to be highly salient to the language model, are by definition the most important sentences for the final autism diagnosis classification. Examples of the attention pattern for every sentence in a particular report are given in Figure 4A, as well as a summary of the key aspects that are mentioned in the most highly attended sentence in each report. Our specialized attention mechanism clearly identified sentences containing specific topics that are traditionally associated with a diagnosis of autism. Notably, the pattern of how the language model spends attentional budget tended to be concentrated, indicating that the model identified a typically small number of very important sentences to highlight for the classification task in a report at hand.

After inspecting the most highlighted sentences from a broader perspective, we identified the most frequently occurring words from the top-attended sentences across all reports. We quantified this by taking the ratio between the number of occurrences of a given word in the most highly attended sentences from diagnosed autism reports and the number of occurrences of that same word in the spotlighted sentences from reports not diagnosed as autism. The most frequently used words that clinicians employed to describe subjects with a final diagnosis of autism compared with those without a diagnosis of autism involved concepts indicative of repetitive movements and speech, special interests, and sensory-processing and perception-based behavior. For example, the word "flapping" occurred 21.5× more often in reports from autism-diagnosed patients compared with reports from children without a diagnosis of autism. Similarly, the words "echolalia" and "vocalizations" occurred 14.1× and 12.2× more often, respectively, in the reports of what turned out to be autism cases. In regard to special interests, the words "letters," "numbers," and "alphabet" were mentioned 24.1×, 16.8×, and 14.0× more frequently in reports of autism cases, respectively.[28] These insights show that our single-sentence neural network approach allowed us

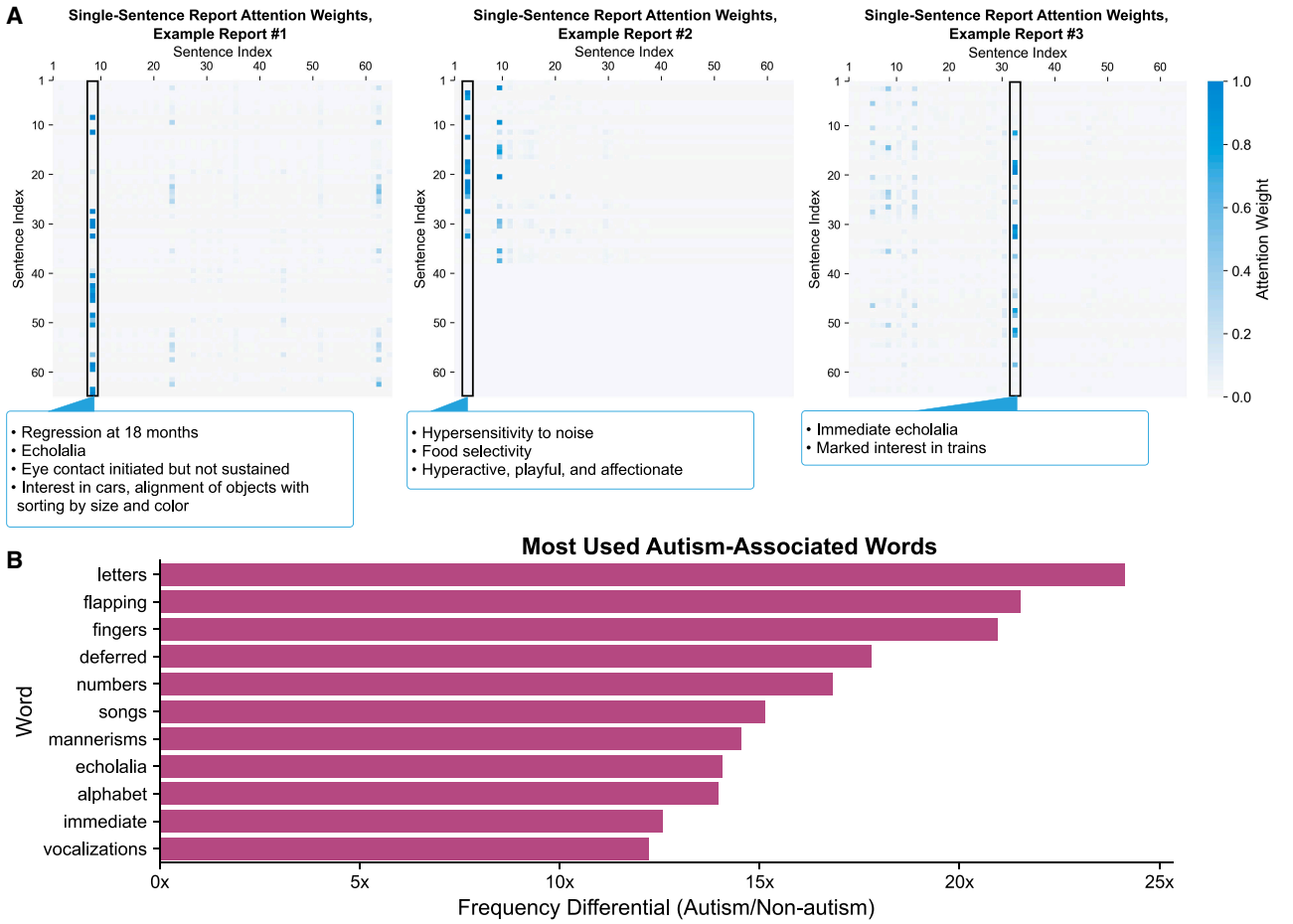


Figure 4. Language model-attention mechanism detects most autism-relevant sentences

(A) Our single-head attention module automatically pinpoints the most essential sentences in a given report for the eventual autism diagnosis classification. Each column/row of the attention weight matrix corresponds to a single sentence for a report, where each cell in the attention matrix corresponds to the semantic link from one sentence on another. Darker color denotes larger attention weight, that is, larger relevance for internal processing in the LLM. The highlighted column indicates the maximally attended sentence in each example report. These classification-driving sentences highlight a diverse array of autism-associated behaviors and developmental histories, summarized to protect patient confidentiality. Each of these example reports was correctly classified as being associated with a clinically confirmed diagnosis of autism.

(B) The single most attended sentences for autistic subjects, per report, contain many times more references to repetitive behaviors and stereotypical behaviors (flapping, "mannerisms," and echolalia), special interests (such as written material: letters, numbers, and alphabet), and verbal/language autistic specificities (vocalizations). The y axis indicates how many times more frequently, in terms of the ratio of raw word occurrences, a given word occurs in the most highly attended sentences in reports from the diagnosed autism cohort compared with those from reports from the cohort that did not receive an autism diagnosis. This word-level breakdown offers a synopsis of the content of our attention-highlighted autism-relevant sentences.

to successfully pinpoint the most important sentences for autism diagnosis and to quantify their importance in relation to all other sentences in each report. These highly up-weighted sentences contain precise terms that are known to be autism-relevant (Figures 4A and 4B).

Transfer learning based on the language model's semantic space allows revisiting the incumbent DSM criteria

To better understand the workings of clinical intuition of practicing healthcare professionals when assessing individuals suspected to carry autism, we developed a modeling tactic that facilitated the inclusion of established diagnostic criteria from the DSM-5 catalog in the context of the embedding space of our language model. Briefly, the DSM-5 catalog contains seven distinct diagnostic criteria for autism, grouped into A and B sections. The classification's A section addresses persistent deficits in social communication and interaction. For instance, A1 denotes "deficits in social-emotional reciprocity," A2 denotes "deficits in nonverbal communicative behaviors used for social interaction," and A3 denotes "deficits in developing, maintaining, and understanding relationships." The B section is dedicated to restricted, repetitive patterns of behavior, interests, or activities. In particular, B1 denotes "stereotyped or repetitive motor movements, use of objects, or speech," B2 denotes "insistence on sameness,

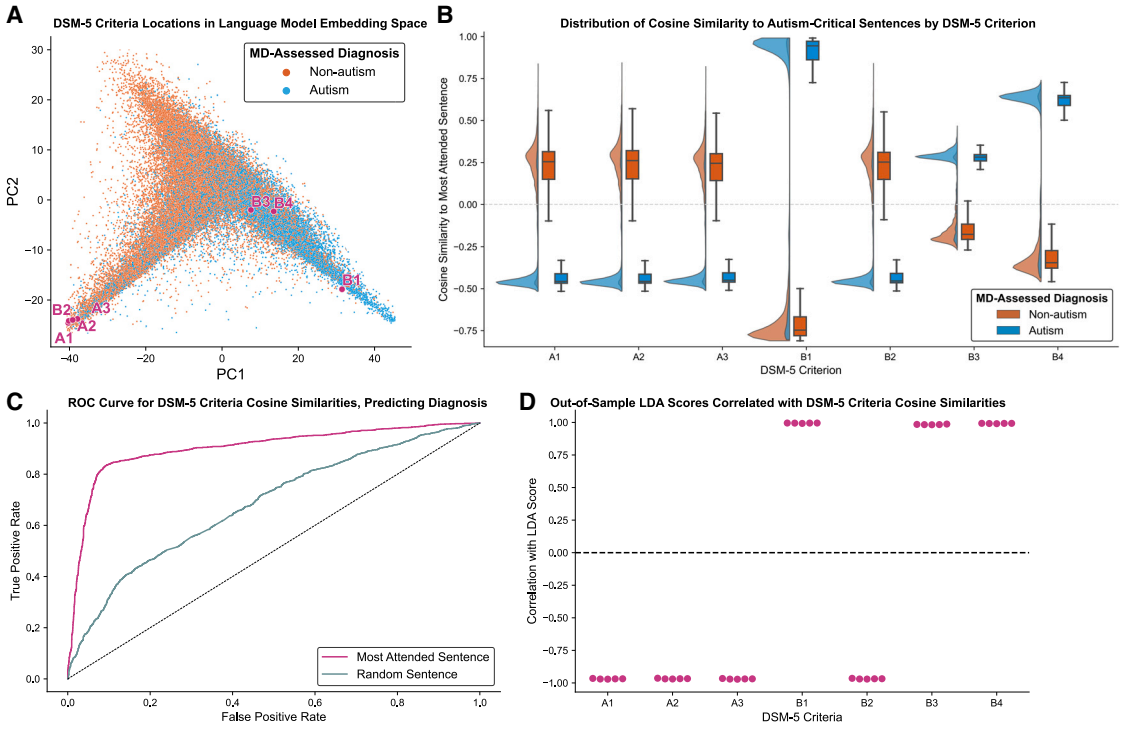


Figure 5. Language model semantic embeddings can be compared against external DSM-5 criteria for autism diagnosis

(A) Using our fine-tuned language model, we can generate meaningful sentence embeddings for any external natural language input sentence. In this case, we generated embeddings for the seven DSM-5 autism criteria—none of which have actually been assessed in our participants by the healthcare professionals. Each criterion can be visualized in the 2D PCA representation space spanned by our language model. B1, B3, and B4 are located closer to the autism-dominant section of the embedding space, whereas A1, A2, A3, and B2 are clustered much more closely together and are situated near the non-autism-dominant section of the embedding space.

(B) After processing each DSM-5 autism criterion (A1–B4) through our model and obtaining sentence-level embeddings, we calculate the cosine similarities between each criterion's embedding and the most attended sentence in each report. This distribution is represented by a density plot as well as a boxplot corresponding to the interquartile range for the cosine similarities between the DSM-5 criteria and the most attended and autism-critical sentence in each report. These distributions are divided based on the final diagnosis for each report (autism versus non-autism). For some criteria, the similarity between that criterion and the most attended sentences follows a very different distribution between autism and non-autism-diagnosed subjects. This suggests that the similarity between DSM-5 criteria embeddings and our model's highly attended sentences can distinguish autism from non-autism subjects.

(C) ROC curve showing the out-of-sample classification performance of the LDA model trained on the cosine similarities of each DSM-5 criterion and the most attended sentences in each report (purple line). Classification performance drops significantly when the cosine similarities of each DSM-5 criterion and a random sentence from each report are used as features for the LDA model (teal line). The derived similarity and dissimilarity indices of the DSM-5 criteria with the top-attended sentence in each report can be used by another machine learning model (LDA) to successfully discriminate between autism and non-autism groups.

(D) Correlation of LDA-transformed scores with DSM-5 criterion cosine similarities is shown for each fold in a 5-fold cross-validation. Each dot represents a single cross-validation fold from the out-of-sample correlation for a criterion, measuring how similar the top-attended sentence for each report is to the DSM-5 diagnostic criteria. LDA scores represent data dimensions that maximize discrimination between autism and non-autism subjects. Correlations indicate the discriminative power of each criterion: values near +1 suggest association with autism, while values near −1 suggest association with non-autism. In particular, B1, B3, and B4 are overall more discriminative for autism in our sample.

See also Figures S2 and S3.

inflexible adherence to routines, or ritualized patterns of verbal or nonverbal behavior," B3 denotes "highly restricted, fixated interests that are abnormal in intensity or focus," and B4 denotes "hyper- or hyporeactivity to sensory input or unusual interests in sensory aspects of the environment." See Table S2 for a full reproduction of the DSM-5 A and B signs for autism. Note that the presence or absence of each DSM-5 criterion was not explicitly listed in the reports.

Nevertheless, we devised a scheme to "rate" each report sentence according to these diagnostic dimensions widespread in clinical practice. Indeed, a major asset of our sentence-specific, language model-based approach (cf. Bzdok et al.[20]) is that it enables the quantitative and objective evaluation and comparison of external descriptors based on localizing their position in our autism-aware semantic embedding space (Figure 5A). Consequently, we strove to evaluate how useful each individual DSM-5 autism criterion is for the diagnosis of autism in the context of the semantic space spanned by our report corpus. We fed all seven DSM-5 criteria descriptions into our model, generating embeddings for each criterion. The A1-A3 and B2 criteria were all positioned in a tight region of the embedding space that was populated mostly by sentences from subjects

with suspected but clinically ruled-out diagnoses of autism. By contrast, the B1, B3, and B4 criteria were more dispersed and located in a region that is highly enriched with autism-associated sentences. These findings align with macro-level trends observed in our word-frequency analysis: clinicians commonly used expressions related to stereotyped movements and speech, special interests, and reactivity to sensory input to describe subjects with confirmed autism, without emphasizing deficits in social communication or interaction. These findings in embedding similarities validated that the overall semantic meaning of the most influential model-identified sentences directly correspond to the DSM-5 criteria regarding stereotyped or repetitive behavior and sensory reactivity.

After embedding each external DSM item in our language model space, we measured the distance in meaning (cosine similarity) between each DSM-5 criterion and the leading autism-related sentence for each report. This analytical protocol yielded seven distinct cosine similarities for each report: cosine similarity is a well-established measure of the semantic similarity between any two instances of sentence embeddings.[29,30] A cosine similarity of +1, 0, or −1 indicates that a given pair of two sentences has identical, no, or opposite semantic meaning, respectively. The cosine similarity group distributions revealed that DSM-5 criteria B1, B3, and B4 are the most similar to the semantic representation of the top sentences distinguishing actual from suspected cases in autism assessment reports. By contrast, the A1-A3 and B2 criteria showed a slightly opposing content of meaning with the top sentences in the diagnosed autism reports (Figure 5B). Further, aside from the B1 criterion, the majority of the top sentences from the suspected but clinically ruled-out autism reports showed near 0 cosine similarity to the DSM-5 criteria, passing an acid test suggesting the soundness of our external validation approach based on the DSM-5 gold standard.

In conjunction with the overlapping positions of A1-A3 and B2 in the embedding space (Figure 5A), our language model interprets these criteria as being somewhat diagnostically homologous and providing relatively less information for distinguishing children with clinically confirmed versus suspected but ruled-out autism. This model-inferred semantic redundancy suggests that each criterion by itself may have reduced diagnostic relevance at an individual level. Taken together, our autism-aware semantic space identified the B1, B3, and B4 criteria as being especially relevant to the confirmatory diagnosis of autism, with the A1-A3 and B2 criteria carrying limited autism-critical information in our cohort. Additional negative controls with an embedding space formed through an orthogonal fine-tuning objective (either random labels or age group prediction) failed to highlight B1, B3, and B4 (see Figures S2 and S3). Thus, B1, B3, and B4 are relevant to the diagnosis outcome specifically and are not simply background features in our dataset.

Finally, we tested whether cosine distances derived from external diagnostic rules were clinically meaningful by using them as input features for a linear discriminant analysis (LDA) classifier. The goal was to detect confirmed autism diagnoses based solely on the semantic similarities between each report's top sentence and the seven DSM-5 criteria. That is, each report, instead of its original text content, was indexed only by seven distinct cosine similarities, representing how closely each report was semantically related to each of the DSM-5 criteria. We were again able to predict the diagnosis in reports previously unseen by the LDA model, achieving a confident AUC of 0.905 (SD, 0.013), averaged across cross-validation folds, in new unseen reports from unseen patients. This observation indicated that these similarity scores do indeed convey compact information that is useful in the context of autism diagnosis (Figure 5C). Moreover, by inspecting the obtained LDA model, we sought to confirm that the B1, B3, and B4 criteria in particular are most predictive in the direction of autism. LDA seeks to calculate a linear combination of input features that allows us to discriminate between the two groups. We found that the DSM sentence similarities that were most helpful in predicting the autism diagnosis for each report were indeed B1, B3, and B4.

As a secondary confirmation that our single-sentence language model-attention approach identifies the most critical sentences for autism classification, we fitted an additional LDA classifier on the cosine similarities between a randomly selected sentence from each report and each of the DSM-5 autism criteria. We found that there was a dramatic drop in the diagnosis classification performance with an average AUC of only 0.676 (SD, 0.021) (Figure 5C). Hence, we confirmed that the level of semantic agreement between the language model-identified influential sentences and the DSM-5 criteria is indeed discriminative between suspected and confirmed autism cases, even on unseen reports. Accordingly, our pattern-learning classifier validated that the B1, B3, and B4 criteria relating to stereotyped or repetitive behavior, special interests, and sensory reactivity are, again, directly predictive of confirmed autism.

DISCUSSION

Some scientists may argue that brain imaging and common-variant genetics in mental health are costly and still not relevant for clinical diagnosis and intervention. These techniques also yielded scarce information on the biological mechanisms that lead to major neurodevelopmental disorders like autism. We argue that the longstanding experience and expertise of health-care professionals working alongside people with autism offer a rich resource to unravel the nature of autism.[31,32] In a field that cannot rely upon biological testing methods, breaking down and analyzing subconscious clinical thought and decision-making processes can potentially shed light on opaque facets of the autism phenotype. In particular, first-hand clinical observation provides an unfiltered verbatim portrait of autism-critical traits and behaviors. In our study, we hence aimed to take the clinical intuition of health professionals itself into sharp focus. For this purpose, we built and deployed a customized language model framework in >4,000 health records on >1,000 children with suspected autism, tailored for single-sentence explainability for direct human interpretation. With these solutions for advanced NLP,[20] we were able to probe and dissect aspects of the diagnostic process that are intrinsically specific to autism in a more impartial way.

For a number of decades, the notion of "specificity" in autism has been a central, and sometimes vexing, topic of dispute.[33] With regards to the description and diagnosis based on DSM-5 criteria, no single clinical autism criterion is pathognomonic.

Many different combinations of signs can lead to the same diagnosis according to the DSM-5 criteria. Some combinations of the same signs may even be better explained by an alternative diagnosis. This study has confronted this challenge in defining autism in a data-driven fashion: our approach carved out the unique facets of behavior, actions, and routine that are most distinctly reflective of autism, in comparison to many conditions and contexts that lead healthcare professionals to initially suspect autism cases. Thanks to the recent acceleration of innovation in the capabilities of language models, combined with our data resource of first-hand clinical observations, we have been able to deconvolve indispensable features consistent across thousands of clinical examinations to directly interrogate pre-existing conceptualizations of autism-specific traits, as codified in gold standard diagnostic manuals such as the DSM-5.

In view of accumulating evidence,[34] there is an urgent need for revisiting the diagnostic criteria that are in default use every day in our mental health institutions. It is instructive to consider that the very term "autism" has been built on shifting ground. "Autism" was initially coined by Bleuler at the beginning of the 20th century to describe severe cases of schizophrenia, with a particular emphasis on avoidance of reality and excessive internal fantasizing.[35,36] By the 1970s, the term autism had undergone a definitional reversal, then referring instead to a lack of internal fantasy and, crucially, a failure to develop social relationships.[37] Building upon these notions, Wing and Gould introduced a system of classification for autism as per childhood impairments anchored in the quality of social interaction, which pushed the idea of autism as a primarily social impairment to the forefront of mainstream autism research.[38]

Indeed, ~40 years ago, an influential study attempted to explain the root of these social deficits. Baron-Cohen, Leslie, and Frith asserted that what is specific to children with autism is a lack of social skills in the form of a "theory of mind" deficit.[39] These authors tested the ability of 27 typically developing children, 14 children with an intellectual disability due to Down syndrome, and 20 children with autism to discern the beliefs held by a hypothetical character in a story. Children with autism, in contrast to children in the two control groups, usually failed to recognize that the story character would hold a false belief about where an object was placed when informed that the object was moved from its original location without the character's knowledge. According to Baron-Cohen and co-investigators, these results illustrated that children with autism are uniquely incapable of inferring the mental states of others—in other words, these children lack a theory of mind. By including children with Down syndrome as a control group, the authors concluded that this intrinsic theory of mind deficit in children with autism is "independent of mental retardation and specific to autism."

The claim of specific inability to conceptualize the mental states of others is highlighted as an underlying cause of social deficits in children with autism—it has served as a North Star for decades of research, diagnosis, and treatment. This conceptualization has had a profound impact on the general understanding of autism and the development of interventions aimed at improving social and communication skills.[40–45] The ever-shifting definition and inconsistent reframing of autism-defining traits witness a long-lasting quest for specificity since the term's inception. Despite the attempts to systematize the characterization of autistic traits in standardized diagnostic instruments, typically taking the form of item checklists, stakeholders are still wondering which criteria are most effective in differentiating individuals with and without autism.[46] Importantly, the emphasis on social communication deficits in today's DSM-5 criteria does align with the theory of mind deficits highlighted by the seminal work of Baron-Cohen and colleagues.

Yet, as a consequence of our collective findings, we call into question the heavy focus on social deficits in research and clinical practice, echoed in established diagnostic instruments that are widely used by clinicians. In contrast to the dominant conceptualization, we found that the most autism-distinctive criteria pertain to repetitive and stereotypical perception-based behaviors, special interests, and sensory reactivity—as predicted by the enhanced perceptual functioning model.[47,48] In our large cohort of suspected and confirmed autism cases, none of the DSM-5 criteria associated with social deficits turned out to be specific to autism. Further, by means of our term frequency analysis of the most autism-predictive sentences, we found that our language model did not identify aspects of social skills as playing an important role in the reports of confirmed autism cases. It should be emphasized that these results are based on the real-world clinical observations of experienced autism practitioners in a clinic that evaluates suspected autism cases on an almost daily basis.

Taken together, our results indicated that, empirically, the most autism-predictive sentences noted by these experienced autism practitioners were highly semantically similar to the DSM-5 criteria relating to repetitive behavior and sensory reactivity. There was essentially no semantic overlap between autism-predictive sentences and DSM-5 criteria regarding social deficits. It may in fact be the case that these repetitive, special interest, and perception-based behaviors are much more prototypical of autism than mainstream research and the clinical state-of-the-art suggest—such a conclusion would be in line with our language model-derived findings.

Despite recently increasing efforts,[49] there is still insufficient emphasis on repetitive behavioral and perception-based traits in standardized diagnostic catalogs. Concretely, following the incumbent DSM-5 manual, the clinician needs to check off three out of three social deficit criteria for a diagnosis of autism. On the other hand, this diagnosis requires only two out of four possible criteria relating to repetitive behaviors, special interests, and sensory reactivity. We argue that this strong weighting on observed features of social deficits in standardized instruments and authoritative medical classification systems is at odds with our language model-enabled quantitative evidence, based on our quantitative assessment of the most autism-predictive words, phrases, and diagnostic criteria employed by experienced clinicians to describe children with autism. In other words, our findings suggest that the social features alone are not specific enough to assign the autism diagnosis to a child. While repetitive and stereotyped behaviors do exist in conditions other than autism,[50,51] it is perhaps their presence in context with sufficiently obvious signs of social deficit that is truly distinctive.

Indeed, as our cohort consisted of suspected autism cases who had been referred to a specialized autism clinic, these

repetitive and stereotyped behaviors are most discriminatory in the direction of autism amidst a shared backdrop of apparent social deficiency. Being more specific and pathognomonic to autism when present alongside socio-communicative deficits, repetitive and perception-based behaviors, and hyperfocused interests is much easier, less time consuming, and more consistent to assess clinically.[52] Moreover, their uniqueness and predictive value have been suspected repeatedly to be superior to highly variable presentations of socio-communicative skills.[53,54] These deficits in socio-communicative skills can exhibit superficial phenotypic similarity in a range of different neurodevelopmental and psychiatric conditions and contexts, such as social (pragmatic) communication disorder, expressive language disorder, ADHD, social anxiety traits, avoidant personality traits, and avoidant attachment style.[55,56] These social behavioral tendencies are also more prone to change over the span of multiple years compared with specific repetitive and perception-based behaviors,[57,58] making these social traits somewhat of a moving target and harder to detect in older children and adults. It may also be said that repetitive and stereotyped behaviors are more salient in a younger cohort representative of first-time autism diagnostic referrals such as ours. These types of behaviors are commonly more prominent in younger children with lower developmental levels, whether diagnosed with autism or not.[59] Hence, due to this clear imbalance in diagnostic criteria, healthcare professionals are spending more time on evaluating traits that, according to our study, have less traction for the overall diagnosis.

From a broader perspective, existing diagnostic tools are framed around what autistic children lack socially and behaviorally, instead of focusing on their cognitive and perception-based strengths and proficiencies. Clearly, the absence of behavior is intrinsically less defined, less idiosyncratic, and less information-dense than the positive signals stemming from a specific behavior. This current framing has widespread implications for potential interventions or behavioral therapeutic strategies. Many current approaches to therapy for autistic individuals focus solely on remediating deficits in social skills and attempt to ameliorate these impairments through direct exposure.[60–62] While these therapeutic interventions are likely to be helpful for many individuals in developing a fulfilling life and improving adaptive functioning, we may see even greater overall quality of life improvements if we focus our energy equally on the cognitive and perception-based strengths and interests of autistic children that are most likely to foster self-esteem and promote positive mental health.[63] As an alternative approach, strength-informed intervention focuses on building on autistic interests and capabilities displayed in natural settings, often reflected in repetitive behaviors and interactions with physical objects.[63]

Given the apparent etiological heterogeneity of autism, it is unclear just how much progress can be expected in identifying specific markers in the general autism population. Indeed, autism has not been described by one definitive biological cause, brain region, gene, or fully consistent set of symptoms. While the biological underpinnings of the disorder may remain diffuse, the clinical presentation of autism may be more uniform,[64] with common threads being shared among differentiated substrata of the autism population. It is such empirical and clinically derived subtyping, perhaps enabled through language model approaches, that could be most helpful for identifying and stratifying individuals that share common underlying neurobiological features, with the goal of introducing more tailored diagnostic criteria or concrete markers.

In conclusion, seizing a newly emerged opportunity, language models hold potential in deconvolving the elements that underpin human intuition at work on the clinical ward. Leveraging this technology can help reconceptualize our understanding and resharpen the focus of our diagnostic tools for many mental health disorders, not just autism.[20] Language models provide a complementary lens, as we have shown here in the example of autism. In today's absence of accurate and objective biomarkers, a challenge pervasive across psychiatry, clinical assessment remains the bedrock upon which major mental health categories are articulated, assessed, and addressed. Ignoring this rich source of actionable insight bears high opportunity costs. We should use every tool at our disposal to put psychiatry on a more solid foundation. Fortunately, text-based AI solutions are becoming more powerful, interpretable, and accessible every day.

Limitations of the study

A few limitations should be borne in mind when considering this study. First, our sample is large and includes subjects from diverse backgrounds. Yet, all subjects were recruited from one region in Northern Montreal, in Québec, Canada. Second, our cohort is relatively young and characteristic of first-time autism referrals. Therefore, the findings presented in this study may be more specific to this age range of autistic subjects and may not be as applicable to older individuals with autism. In general, we present our results in the context of the full widely sampled cohort. That is, we do not attempt to draw more narrow conclusions on the basis of sex, specific age groups, or other demographic variables. Investigations into these demographic subsets are a topic for future work. Finally, our findings are focused on traits unique to autism. That is, the clinician-noted behaviors that we extract are differentiated between those who are clinically confirmed to have autism and those who are suspected to have autism but have been clinically ruled out, not neurotypicals.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Danilo Bzdok (danilo.bzdok@mcgill.ca).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The clinical text data and associated autism diagnoses reported in this study cannot be deposited in a public repository because they contain highly identifiable and sensitive information regarding children. To request access, please contact the Autism Spectrum Disorder Assessment Clinic at CIUSSS-Nord-de-l'île-de-Montréal. The primary point of contact for data access is Laurent Mottron (laurent.mottron.cnmtl@ssss.gouv.qc.ca).

- All original code has been deposited at https://github.com/dblabs-mcgill-mila/NLP-ASD and is publicly available and archived at https://doi.org/10.5281/zenodo.14851367 as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

D.B. was supported by the Brain Canada Foundation, through the Canada Brain Research Fund, with the financial support of Health Canada, National Institutes of Health (NIH R01 AG068563A, NIH R01 DA053301-01A1, and NIH R01 MH129858-01A1), the Canadian Institute of Health Research (CIHR 438531 and CIHR 470425), the Healthy Brains Healthy Lives initiative (Canada First Research Excellence fund), the IVADO R3AI initiative (Canada First Research Excellence fund), and the CIFAR Artificial Intelligence Chairs program (Canada Institute for Advanced Research). Icons used in the graphical abstract and Figure 1 were provided by BioRender.

AUTHOR CONTRIBUTIONS

L.M. and D.B. conceived and initiated the study. E.R. and L.M. ensured ethical handling of the data (including manual censoring of sensitive person-specific information) and contributed proper interpretation and contextualization of results based on clinical expertise in autism diagnosis. E.R. was responsible for coordination and administration of clinical data collection. J.S. and D.B. conceptualized the quantitative modeling approach, developed interpretable language model architectures, and carried out all subsequent data analysis steps. J.S., E.R., L.M., and D.B. wrote the manuscript, with critical revisions and modifications from S.R. and E.B. D.B. led data analysis.

DECLARATION OF INTERESTS

D.B. is a shareholder and advisory board member at MindState Design Labs, USA, as well as a shareholder of Biossil.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Participant recruitment
  - Clinical context
  - Inclusion and exclusion criteria
- METHOD DETAILS
  - Assessment format
  - Data format
  - Data processing
  - Pre-trained language model
  - Language model architecture enhancements
  - Fine-tuning framework to spawn an autism-aware embedding representation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Tactics tailored for model interpretability
  - Evaluation against external diagnostic criteria

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2025.02.025.

Received: May 24, 2024
Revised: October 28, 2024
Accepted: February 21, 2025
Published: March 26, 2025

REFERENCES

1. Xu, G., Strathearn, L., Liu, B., and Bao, W. (2018). Prevalence of Autism Spectrum Disorder Among US Children and Adolescents, 2014-2016. JAMA *319*, 81–82. https://doi.org/10.1001/jama.2017.17812.
2. Poldrack, R.A., and Gorgolewski, K.J. (2014). Making big data open: data sharing in neuroimaging. Nat. Neurosci. *17*, 1510–1517. https://doi.org/10.1038/nn.3818.
3. van den Heuvel, M.P., and Hulshoff Pol, H.E. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. Eur. Neuropsychopharmacol. *20*, 519–534. https://doi.org/10.1016/j.euroneuro.2010.03.008.
4. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.
5. Schaefer, G.B., and Mendelsohn, N.J. (2008). Genetics evaluation for the etiologic diagnosis of autism spectrum disorders. Genet. Med. *10*, 4–12. https://doi.org/10.1097/GIM.0b013e31815efdd7.
6. Tick, B., Bolton, P., Happé, F., Rutter, M., and Rijsdijk, F. (2016). Heritability of autism spectrum disorders: a meta-analysis of twin studies. J. Child Psychol. Psychiatry *57*, 585–595. https://doi.org/10.1111/jcpp.12499.
7. Antaki, D., Guevara, J., Maihofer, A.X., Klein, M., Gujral, M., Grove, J., Carey, C.E., Hong, O., Arranz, M.J., Hervas, A., et al. (2022). A phenotypic spectrum of autism is attributable to the combined effects of rare variants, polygenic risk and sex. Nat. Genet. *54*, 1284–1292. https://doi.org/10.1038/s41588-022-01064-5.
8. Wray, N.R., Lin, T., Austin, J., McGrath, J.J., Hickie, I.B., Murray, G.K., and Visscher, P.M. (2021). From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. JAMA Psychiatry *78*, 101–109. https://doi.org/10.1001/jamapsychiatry.2020.3049.
9. Grove, J., Ripke, S., Als, T.D., Mattheisen, M., Walters, R.K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O.A., Anney, R., et al. (2019). Identification of common genetic risk variants for autism spectrum disorder. Nat. Genet. *51*, 431–444. https://doi.org/10.1038/s41588-019-0344-8.
10. Skafidas, E., Testa, R., Zantomio, D., Chana, G., Everall, I.P., and Pantelis, C. (2014). Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. Mol. Psychiatry *19*, 504–510. https://doi.org/10.1038/mp.2012.126.
11. Myers, S.M., Challman, T.D., Bernier, R., Bourgeron, T., Chung, W.K., Constantino, J.N., Eichler, E.E., Jacquemont, S., Miller, D.T., Mitchell, K.J., et al. (2020). Insufficient Evidence for "Autism-Specific" Genes. Am. J. Hum. Genet. *106*, 587–595. https://doi.org/10.1016/j.ajhg.2020.04.004.
12. Iakoucheva, L.M., Muotri, A.R., and Sebat, J. (2019). Getting to the Cores of Autism. Cell *178*, 1287–1298. https://doi.org/10.1016/j.cell.2019.07.037.
13. Geschwind, D.H. (2008). Autism: Many Genes, Common Pathways? Cell *135*, 391–395. https://doi.org/10.1016/j.cell.2008.10.016.
14. Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D., Thirion, B., and Varoquaux, G. (2017). Deriving reproducible biomarkers from multi-site resting-state data: an Autism-based example. NeuroImage *147*, 736–745. https://doi.org/10.1016/j.neuroimage.2016.10.045.
15. Nielsen, J.A., Zielinski, B.A., Fletcher, P.T., Alexander, A.L., Lange, N., Bigler, E.D., Lainhart, J.E., and Anderson, J.S. (2013). Multisite functional connectivity MRI classification of autism: ABIDE results. Front. Hum. Neurosci. *7*, 599. https://doi.org/10.3389/fnhum.2013.00599.
16. Heinsfeld, A.S., Franco, A.R., Craddock, R.C., Buchweitz, A., and Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. NeuroImage Clin. *17*, 16–23. https://doi.org/10.1016/j.nicl.2017.08.017.

17. Poulin-Lord, M.-P., Barbeau, E.B., Soulières, I., Monchi, O., Doyon, J., Benali, H., and Mottron, L. (2014). Increased topographical variability of task-related activation in perceptive and motor associative regions in adult autistics. NeuroImage Clin. *4*, 444–453. https://doi.org/10.1016/j.nicl.2014.02.008.

18. Weinberger, D.R., and Radulescu, E. (2016). Finding the Elusive Psychiatric "Lesion" With 21st-Century Neuroanatomy: A Note of Caution. Am. J. Psychiatry *173*, 27–33. https://doi.org/10.1176/appi.ajp.2015.15060753.

19. Arvidsson, O., Gillberg, C., Lichtenstein, P., and Lundström, S. (2018). Secular changes in the symptom level of clinically diagnosed autism. J. Child Psychol. Psychiatry *59*, 744–751. https://doi.org/10.1111/jcpp.12864.

20. Bzdok, D., Thieme, A., Levkovskyy, O., Wren, P., Ray, T., and Reddy, S. (2024). Data science opportunities of large language models for neuroscience and biomedicine. Neuron *112*, 698–717. https://doi.org/10.1016/j.neuron.2024.01.016.

21. American Psychiatric Association (2013). Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (American Psychiatric Association) https://doi.org/10.1176/appi.books.9780890425596.

22. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. Preprint at arXiv. https://doi.org/10.48550/arXiv.1907.11692.

23. Le, Q., and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning (PMLR), pp. 1188–1196.

24. Beltagy, I., Peters, M.E., and Cohan, A. (2020). Longformer: The Long-Document Transformer. Preprint at arXiv. https://doi.org/10.48550/arXiv.2004.05150.

25. Bazoge, A., Morin, E., Daille, B., and Gourraud, P.-A. (2024). Adaptation of Biomedical and Clinical Pretrained Models to French Long Documents: A Comparative Study. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.16689.

26. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The Llama 3 Herd of Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2407.21783.

27. Gemma Team, Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., and Love, J. (2024). Gemma: Open Models Based on Gemini Research and Technology. Preprint at arXiv. https://doi.org/10.48550/arXiv.2403.08295.

28. Ostrolenk, A., Gagnon, D., Boisvert, M., Lemire, O., Dick, S.-C., Côté, M.-P., and Mottron, L. (2024). Enhanced interest in letters and numbers in autistic children. Mol. Autism *15*, 26. https://doi.org/10.1186/s13229-024-00606-4.

29. Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Preprint at arXiv. https://doi.org/10.48550/arXiv.1301.3781.

30. Reimers, N., and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1908.10084.

31. Mottron, L. (2021). Progress in autism research requires several recognition-definition-investigation cycles. Autism Res. *14*, 2230–2234. https://doi.org/10.1002/aur.2524.

32. Mottron, L. (2011). Changing perceptions: The power of autism. Nature *479*, 33–35. https://doi.org/10.1038/479033a.

33. Mottron, L. (2021). A radical change in our autism research strategy is needed: Back to prototypes. Autism Res. *14*, 2213–2220. https://doi.org/10.1002/aur.2494.

34. Bzdok, D., and Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. Biol. Psychiatry Cogn. Neurosci. Neuroimaging *3*, 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007.

35. Bleuler, E. (1950). Dementia Praecox or the Group of Schizophrenias (International Universities Press).

36. Evans, B. (2013). How autism became autism: The radical transformation of a central concept of child development in Britain. Hist. Human Sci. *26*, 3–31. https://doi.org/10.1177/0952695113484320.

37. Rutter, M. (1972). Childhood schizophrenia reconsidered. J. Autism Child. Schizophr. *2*, 315–337. https://doi.org/10.1007/BF01537622.

38. Wing, L., and Gould, J. (1979). Severe impairments of social interaction and associated abnormalities in children: Epidemiology and classification. J. Autism Dev. Disord. *9*, 11–29. https://doi.org/10.1007/BF01531288.

39. Baron-Cohen, S., Leslie, A.M., and Frith, U. (1985). Does the autistic child have a "theory of mind"? Cognition *21*, 37–46. https://doi.org/10.1016/0010-0277(85)90022-8.

40. Frith, U. (1989). Autism and "Theory of Mind". In Diagnosis and Treatment of Autism, C. Gillberg, ed. (Springer), pp. 33–52. https://doi.org/10.1007/978-1-4899-0882-7_4.

41. Charman, T. (2003). Why is joint attention a pivotal skill in autism? Philos. Trans. R. Soc. Lond. B Biol. Sci. *358*, 315–324. https://doi.org/10.1098/rstb.2002.1199.

42. Happé, F.G.E., and Frith, U. (1996). Theory of mind and social impairment in children with conduct disorder. Br. J. Dev. Psychol. *14*, 385–398. https://doi.org/10.1111/j.2044-835X.1996.tb00713.x.

43. Mundy, P., and Crowson, M. (1997). Joint Attention and Early Social Communication: Implications for Research on Intervention with Autism. J. Autism Dev. Disord. *27*, 653–676. https://doi.org/10.1023/A:1025802832021.

44. Sigman, M. (1998). The Emanuel Miller Memorial Lecture 1997. Change and continuity in the development of children with autism. J. Child Psychol. Psychiatry *39*, 817–827. https://doi.org/10.1111/1469-7610.00383.

45. Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognitiondiscussion 691–735. Behav. Brain Sci. *28*, 675–691. https://doi.org/10.1017/S0140525X05000129.

46. Molloy, C.A., Murray, D.S., Akers, R., Mitchell, T., and Manning-Courtney, P. (2011). Use of the Autism Diagnostic Observation Schedule (ADOS) in a clinical setting. Autism *15*, 143–162. https://doi.org/10.1177/1362361310379241.

47. Mottron, L., Dawson, M., Soulières, I., Hubert, B., and Burack, J. (2006). Enhanced perceptual functioning in autism: an update, and eight principles of autistic perception. J. Autism Dev. Disord. *36*, 27–43. https://doi.org/10.1007/s10803-005-0040-7.

48. Mottron, L., and Gagnon, D. (2023). Prototypical autism: New diagnostic criteria and asymmetrical bifurcation model. Acta Psychol. (Amst.) *237*, 103938. https://doi.org/10.1016/j.actpsy.2023.103938.

49. Uljarević, M., Alvares, G.A., Steele, M., Edwards, J., Frazier, T.W., Hardan, A.Y., and Whitehouse, A.J. (2022). Toward better characterization of restricted and unusual interests in youth with autism. Autism *26*, 1296–1304. https://doi.org/10.1177/13623613211056720.

50. Barnard-Brak, L., Rojahn, J., Richman, D.M., Chesnut, S.R., and Wei, T. (2015). Stereotyped behaviors predicting self-injurious behavior in individuals with intellectual disabilities. Res. Dev. Disabil. *36C*, 419–427. https://doi.org/10.1016/j.ridd.2014.08.017.

51. Brunetti, S., Rossi, A., Galli, J., Gitti, F., Nardocci, N., Giordano, L., Accorsi, P., Calza, S., and Fazzi, E. (2022). Repetitive and stereotyped behaviors in neurodevelopmental disorders: an observational analysis of four diagnostic groups. Published online May 5, 2022. Minerva Pediatr. (Torino). https://doi.org/10.23736/S2724-5276.22.06835-5.

52. Uljarević, M., Jo, B., Frazier, T.W., Scahill, L., Youngstrom, E.A., and Hardan, A.Y. (2021). Using the big data approach to clarify the structure of restricted and repetitive behaviors across the most commonly used autism spectrum disorder measures. Mol. Autism *12*, 39. https://doi.org/10.1186/s13229-021-00419-9.

53. Mottron, L., Mineau, S., Martel, G., Bernier, C.S.-C., Berthiaume, C., Dawson, M., Lemay, M., Palardy, S., Charman, T., and Faubert, J. (2007). Lateral glances toward moving stimuli among young children with autism: early regulation of locally oriented perception? Dev. Psychopathol. *19*, 23–36. https://doi.org/10.1017/S0954579407070022.

54. Ozonoff, S., Macari, S., Young, G.S., Goldring, S., Thompson, M., and Rogers, S.J. (2008). Atypical object exploration at 12 months of age is associated with autism in a prospective sample. Autism *12*, 457–472. https://doi.org/10.1177/1362361308096402.

55. Kamio, Y., Inada, N., Moriwaki, A., Kuroda, M., Koyama, T., Tsujii, H., Kawakubo, Y., Kuwabara, H., Tsuchiya, K.J., Uno, Y., et al. (2013). Quantitative autistic traits ascertained in a national survey of 22 529 Japanese schoolchildren. Acta Psychiatr. Scand. *128*, 45–53. https://doi.org/10.1111/acps.12034.

56. Schilbach, L. (2016). Towards a second-person neuropsychiatry. Philos. Trans. R. Soc. Lond. B Biol. Sci. *371*, 20150081. https://doi.org/10.1098/rstb.2015.0081.

57. Elison, J.T., Wolff, J.J., Reznick, J.S., Botteron, K.N., Estes, A.M., Gu, H., Hazlett, H.C., Meadows, A.J., Paterson, S.J., Zwaigenbaum, L., et al. (2014). Repetitive Behavior in 12-Month-Olds Later Classified With Autism Spectrum Disorder. J. Am. Acad. Child Adolesc. Psychiatry *53*, 1216–1224. https://doi.org/10.1016/j.jaac.2014.08.004.

58. Miller, M., Sun, S., Iosif, A.-M., Young, G.S., Belding, A., Tubbs, A., and Ozonoff, S. (2021). Repetitive behavior with objects in infants developing autism predicts diagnosis and later social behavior as early as 9 months. J. Abnorm. Psychol. *130*, 665–675. https://doi.org/10.1037/abn0000692.

59. Courchesne, V., Bedford, R., Pickles, A., Duku, E., Kerns, C., Mirenda, P., Bennett, T., Georgiades, S., Smith, I.M., Ungar, W.J., et al. (2021). Nonverbal IQ and change in restricted and repetitive behavior throughout childhood in autism: a longitudinal study using the Autism Diagnostic Interview-Revised. Mol. Autism *12*, 57. https://doi.org/10.1186/s13229-021-00461-7.

60. Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., Donaldson, A., and Varley, J. (2010). Randomized, Controlled Trial of an Intervention for Toddlers With Autism: The Early Start Denver Model. Pediatrics *125*, e17–e23. https://doi.org/10.1542/peds.2009-0958.

61. Pickles, A., Le Couteur, A.L., Leadbitter, K., Salomone, E., Cole-Fletcher, R., Tobin, H., Gammer, I., Lowry, J., Vamvakas, G., Byford, S., et al. (2016). Parent-mediated social communication therapy for young children with autism (PACT): long-term follow-up of a randomised controlled trial. Lancet *388*, 2501–2509. https://doi.org/10.1016/S0140-6736(16)31229-6.

62. Shih, W., Shire, S., Chang, Y.-C., and Kasari, C. (2021). Joint engagement is a potential mechanism leading to increased initiations of joint attention and downstream effects on language: JASPER early intervention for children with ASD. J. Child Psychol. Psychiatry *62*, 1228–1235. https://doi.org/10.1111/jcpp.13405.

63. Mottron, L. (2017). Should we change targets and methods of early intervention in autism, in favor of a strengths-based education? Eur. Child Adolesc. Psychiatry *26*, 815–825. https://doi.org/10.1007/s00787-017-0955-5.

64. Mottron, L., and Bzdok, D. (2020). Autism spectrum heterogeneity: fact or artifact? Mol. Psychiatry *25*, 3178–3185. https://doi.org/10.1038/s41380-020-0748-y.

65. van Rossum, G. (2003). The Python Language Reference Manual (Network Theory Limited).

66. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in PyTorch. In 31st Conference on Neural Information Processing Systems (NIPS 2017). https://openreview.net/pdf?id=BJJsrmfCZ.

67. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. Preprint at arXiv. https://doi.org/10.48550/arXiv.1910.03771.

68. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

69. McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

70. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

71. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55.

72. Waskom, M.L. (2021). seaborn: statistical data visualization. J. Open Source Softw. *6*, 3021. https://doi.org/10.21105/joss.03021.

73. Lord, C., Rutter, M., DiLavore, P.C., Risi, S., and Gotham, K. (2012). Autism Diagnostic Observation Schedule, Second Edition (Western Psychological Services).

74. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems (Curran Associates, Inc.), pp. 1877–1901.

75. Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT: Unsupervised Language Model Pre-training for French. Preprint at arXiv.

76. Koehn, P., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, S. Ananiadou, ed. (Association for Computational Linguistics), pp. 177–180. https://doi.org/10.3115/1557769.1557821.

77. Vajre, V., Naylor, M., Kamath, U., and Shehu, A. (2021). PsychBERT: A Mental Health Language Model for Social Media Mental Health Behavioral Analysis. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1077–1082. https://doi.org/10.1109/BIBM52615.2021.9669469.

78. Huang, K., Altosaar, J., and Ranganath, R. (2020). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. Preprint at arXiv. https://doi.org/10.48550/arXiv.1904.05342.

79. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics *36*, 1234–1240. https://doi.org/10.1093/bioinformatics/btz682.

80. Abnar, S., and Zuidema, W. (2020). Quantifying Attention Flow in Transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational (Association for Computational Linguistics), pp. 4190–4197. https://doi.org/10.18653/v1/2020.acl-main.385.

81. Hao, Y., Dong, L., Wei, F., and Xu, K. (2021). Self-Attention Attribution: Interpreting Information Interactions Inside Transformer. In The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), *35*, pp. 12963–12971. https://doi.org/10.1609/aaai.v35i14.17533.

82. Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. (2019). Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, A. Korhonen, D. Traum, and L. Màrquez, eds. (Association for Computational Linguistics), pp. 5797–5808. https://doi.org/10.18653/v1/P19-1580.

83. Bahdanau, D., Cho, K., and Bengio, Y. (2016). Neural Machine Translation by Jointly Learning to Align and Translate. Preprint at arXiv. https://doi.org/10.48550/arXiv.1409.0473.

84. Graves, A., Wayne, G., and Danihelka, I. (2014). Neural Turing Machines. Preprint at arXiv. https://doi.org/10.48550/arXiv.1410.5401.

85. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. In Advances in Neural Information Processing Systems (Curran Associates, Inc.).

86. Kingma, D.P., and Ba, J. (2017). Adam: A Method for Stochastic Optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6980.

87. Loshchilov, I., and Hutter, F. (2019). Decoupled Weight Decay Regularization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1711.05101.

88. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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), J. Burstein, C. Doran, and T. Solorio, eds. (Association for Computational Linguistics), pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

89. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2106.09685.

90. Biderman, D., Portes, J., Ortiz, J.J.G., Paul, M., Greengard, P., Jennings, C., King, D., Havens, S., Chiley, V., Frankle, J., et al. (2024). LoRA Learns Less and Forgets Less. Preprint at arXiv. https://doi.org/10.48550/arXiv.2405.09673.

91. Kuhn, M., and Johnson, K. (2013). Discriminant Analysis and Other Linear Classification Models. In Applied Predictive Modeling, M. Kuhn and K. Johnson, eds. (Springer), pp. 275–328. https://doi.org/10.1007/978-1-4614-6849-3_12.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Custom language model and analysis pipeline	This paper	https://doi.org/10.5281/zenodo.14851367
Python 3.10.4	van Rossum[65]	https://www.python.org/downloads/release/python-3104/
PyTorch 2.4.0	Paszke et al.[66]	https://pypi.org/project/torch/
Transformers 4.44.2	Wolf et al.[67]	https://pypi.org/project/transformers/
NumPy 1.26.4	Harris et al.[68]	https://pypi.org/project/numpy/
Pandas 2.2.2	McKinney[69]	https://pypi.org/project/pandas/
scikit-learn 1.5.2	Pedregosa et al.[70]	https://pypi.org/project/scikit-learn/
Matplotlib 3.5.2	Hunter[71]	https://pypi.org/project/matplotlib/
Seaborn 0.11.0	Waskom[72]	https://pypi.org/project/seaborn/
Document AI OCR	Google Cloud Platform	https://cloud.google.com/python/docs/reference/documentai/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Participant recruitment

In the Québec public healthcare system, referrals for an autism spectrum disorder diagnosis for children are typically initiated by a general practitioner or pediatrician. Preschool-aged children are first assessed by professionals within the local early intervention care program, such as psycho-educators, specialized educators, and social workers. These professionals conduct screenings for signs of autism in conjunction with the medical referral.

Clinical context

The Autism Spectrum Disorder Assessment Clinic at CIUSSS-Nord-de-l'île-de-Montréal (CIUSSS-NIM), Rivière-des-Prairies Hospital serves as the sole public healthcare entry point for children referred for an autism spectrum disorder diagnosis within a specifically defined geographic area in Northern Montréal, Québec, Canada. This region experiences approximately 13,500 births annually. This clinic benefits from a concentration of expertise not found elsewhere in Québec. However, children presenting severe neurodevelopmental issues during infancy and early childhood (such as significant motor delay, severe global developmental delay, epilepsy, clinical indicators of a genetic condition) are directed to a pediatric facility for assessment. Therefore, our participant sample approximately represents a cumulative four-year incidence of children exhibiting apparent symptoms of autism warranting assessment, without obvious detectable neurogenetic indicators.

Inclusion and exclusion criteria

1180 patients referred to the Autism Spectrum Disorder Assessment Clinic who received an autism spectrum disorder assessment between January 1st, 2016 to December 31st 2019 were initially screened. We excluded 57 patients (4.8%) that did not have at least two reports written by two different professionals, containing at least one full page of qualitative clinical description, either referral reports and/or assessment reports. We excluded 43 patients (3.6%) for whom the presence of autism was still undetermined or uncertain, meaning that the final diagnosis was not reached by the end of year 2019. This resulted in a total inclusion of 4272 reports corresponding to 1080 participants. Of these 1080 participants, 429 received a diagnosis of autism while 651 did not receive a diagnosis of autism. The age of each participant was recorded at the time of each report; the average age of participants in our cohort was 7.0 years (standard deviation 3.9 years). See Table S3 for a more detailed breakdown of sex and age groups included in our cohort, and see Table S4 for a tabulation of secondary diagnoses present. Tables S5–S7 provide additional accuracy results for our model, stratified on the basis of these variables. For some demographic groups we see slightly degraded model prediction performance. This is likely to be attributable to a smaller number of training examples present in these strata, the inherent diagnostic complexity of these groups, or a combination of the two. For instance, this combination of factors is evidently present for the age bracket consisting of subjects over 12 years of age, resulting in slightly lower classification accuracy compared to other more populated age brackets in our cohort.

This clinical research project, number 2021-2058, was approved by the CIUSSS-NIM ethical committee on May 1st, 2020 and by the CIUSSS-NIM direction of professional services on November 17th, 2020.

METHOD DETAILS

Assessment format

The diagnostic assessment procedure at the Autism Spectrum Disorder Assessment Clinic is carried out by child and adolescent psychiatrists, along with, depending on the clinical situation, a psychologist, a neuropsychologist, a speech therapist, an occupational therapist, and/or a psycho-educator. The evaluating professionals and the clinicians have between 5 to 20 years of experience in autism assessments. Patients are assigned to clinicians randomly, and combinations of psychiatrists and professionals are shuffled. The assessment procedure includes an in-depth developmental and psychiatric interview with the parents and the child, lasting one to two hours, along with behavioral observation of the child. A standardized Autism Diagnostic Observation Schedule (ADOS-2)[73] was typically performed, but may be skipped based on the extreme obviousness of the autism diagnosis or its absence following the clinical interview and non-standardized observation. Supplementary assessments in psychology, neuropsychology, occupational therapy, and speech therapy may be conducted. Interviews with daycare or school staff and/or observation of the child in their facility is added when necessary to gather behavioral data from more ecological settings. Appointments are mainly conducted in French, sometimes in English, or in French with the assistance of an interpreter. All ensuing assessment reports are written in French.

The diagnosis is established by consensus within the multidisciplinary team based on DSM-5 criteria. These diagnostic teams are shuffled for every individual case. Clinical certainty is given priority over the child's ADOS-2 score in cases of discrepancy.[46] Reports are written independently by psychiatrists and professionals. At the end of the assessment process, patients either receive a final diagnosis of autism spectrum disorder, are cleared of autism with no diagnosis, or receive a differential neurodevelopmental or psychiatric diagnosis. It is this final yes/no autism diagnosis that we used as our ground-truth target variable (classification outcome) for language model fine-tuning and subsequent interpretation. It is important to note that this final yes/no diagnosis is not contained anywhere within the reports themselves. In fact, the clinician writing the report has no knowledge of the future diagnosis at the time of writing. The reports serve solely as a dispassionate account of observed behavior of the child or developmental history provided by the parents, to be referenced by the diagnostic team at the time of consensus diagnosis.

Henceforth, we use the terminology "clinically confirmed" and "suspected but clinically ruled-out" to refer to the definitive yes/no autism diagnosis decision as agreed upon by the aforementioned multidisciplinary team of healthcare professionals. Since each subject who is referred to the clinic carries some suspicion of having autism, it is the task of the clinicians to definitively "confirm" or "rule out" this suspicion. While diagnoses reported here represent a best estimate clinical assessment, these methods remain the gold standard for diagnosis in autism.

Data format

Our data consists of raw qualitative clinical descriptions obtained directly from referral and assessment reports. These reports are varied in formatting and content. However, they frequently contain sections relating to the reason for the evaluation, developmental history as provided by parents, first-hand observations from the clinician upon interacting with the child, and miscellaneous notes that may be of diagnostic interest, among others.

Referral Reports

The referral reports are authored by frontline referring professionals (33% of the total included reports) during their clinical routine evaluation of the children and intended for the multidisciplinary assessment team. These professionals include general practitioners, pediatricians, and early intervention program specialists (psycho-educators, specialized educators, social workers, speech therapists). The professionals who refer children to the Autism Spectrum Disorder Assessment Clinic possess extensive exposure to neurodevelopmental presentations and possess strong clinical and developmental expertise. Importantly, information contained in referral reports is collated from several dozen different individuals, diminishing the risk of individual biases.

Assessment Reports

The assessment reports (67% of the total included reports) from the Autism Spectrum Disorder Assessment Clinic were articulated by professionals who are experts in autism within their respective fields, including 10 psychiatrists who practiced at the clinic between 2016 and 2019 (44% of the included assessment reports were from psychiatrists) and other professionals: psychologists, neuropsychologists, speech therapists, occupational therapists, and psycho-educators (a total of 56% of the included assessment reports were from other professionals). The reports were written at the end of the autism spectrum disorder assessment of the children as a detailed report intended for the patients and their referral doctors.

Data Ingestion and Anonymization

The reports were written in French by the healthcare professionals, typed on computers, printed, and then scanned into the hospital's medical software, prior to the study. Following a strict anonymization procedure carried out on-premises at CIUSSS-NIM by the clinical research team, strictly only qualitative descriptions were extracted from the reports. The clinical team finalized these redacted reports in PDF format. All standardized clinical information (e.g., from the ADOS-2 test, standardized tests in psychology, neuropsychology, and speech therapy) was manually excluded by research assistants and verified by a clinician-researcher. Then, the data analysis team, who never received access to the unredacted reports, stored these de-identified reports in secure computing facilities locally at McGill University Brain Imaging Center (BIC), as well as at Mila Québec AI Institute. Due to the sensitivity of this medical information regarding children, we are unable to publicly release verbatim excerpts from the reports. In lieu of these verbatim

excerpts, we have summarized examples of highly autism relevant sentences (Figure 4A), as well as carried out numerous statistical analyses to quantify the semantic content of every report in our dataset (Figures 4B and 5A–5D).

Data processing

Each of the 4272 health records from 1080 unique patients were scanned and converted to digital PDF format. While each report was originally typed out on a computer by a healthcare professional, these reports were printed in hard copy with a variety of different, non-standardized document formats. This meant that the physical placement of text on the page and the location of key pieces of information, as well as formatting of the text, such as font size and type, could be highly variable from report to report. To faithfully extract the raw text, ignoring differences in document formatting and visual artifacts, we turned to state-of-the-art optical character recognition (OCR) from the Google Cloud Platform "Document AI" API. No information was stored on Google's servers. This advanced tool for OCR computer vision ensured that we could automatically convert the text contained within our complex PDF files to simple text files with high fidelity. As a preliminary step, we cleaned the OCR-extracted text files by removing repetitive special characters (such as asterisks, dashes, dollar signs, or tildes) that simply represented minor anomalies in the OCR process. We did not perform any additional steps for text cleaning, such as converting to lowercase or removing numbers or stop words, since it has been noted that pre-trained language models perform optimally even when given text in its most raw format.[74] These minimally cleaned text files could then be easily ingested by our language model as input data.

With these text files in hand for each report, we next wanted to segment each report into individual sentences, to enable straightforward interpretation and identification of key semantic elements contained within each report. To accomplish this sentence-level segmentation, we relied on a straightforward rule: we would split each sentence at a period, unless the resulting sentence was 30 characters in length or less, in which case that shorter sentence stub was added to the previous sentence. This ensured that each sentence contained a sufficient amount of meaningful semantic content as a solid starting point for our language model to analyze. This pre-processing step applied to a very small minority of sentences, of which the vast majority consisted of titles, section headings, or erroneous punctuation introduced by OCR artifacts. No sentences were discarded as a result of this pre-processing. Each report contained a median of 1196 words and 68 distinct sentences. Each sentence contained a median of 15 words, across all reports.

Pre-trained language model

We chose the "Hugging Face" Transformer[67] library to deploy tokenizers and model weights for our pre-trained language model. There were numerous options for our selection of pre-trained language model; critically, we required a transformer encoder architecture that had the capacity and pre-training exposure necessary to effectively parse a wide range of diverse natural language inputs. Ultimately, we selected the FlauBERT language model[75] that was pre-trained on a total of 12.8 billion word tokens from general language sources, amounting to a total of ∼489 million unique sentences.

FlauBERT is a French-language adaptation of the widely utilized RoBERTa architecture, sharing the same underlying model architecture. The key distinction lies in FlauBERT's pre-training: it incorporates a broader corpus of French text, in addition to the internet-scale (primarily English) data sources used to pre-train RoBERTa. We will henceforth refer to our pre-trained model selection as "RoBERTa", as this model is more readily recognized amongst the broader NLP community. In total, the RoBERTa language model contains 138 million distinct parameters, which were all pre-trained in concert on the aforementioned training corpus based on a masked language modeling objective (cf. below). These 138 million parameters were spread across 12 layers, with 12 independent self-attention heads per layer. Individual words were "tokenized" according to the established Moses tokenizer[76] and optimized for the French language (as our clinical reports). These tokenized words could then be fed into the language model and projected into increasingly relevant "embedding" representations as their associated information flowed through the model layers. The term "embedding" refers to numerical vector representations of semantic elements such as words, sentences, or even full reports, which are given meaning and are interpretable by our language model.

Pre-training was accomplished with the masked language modeling objective, wherein 15% of the words in an input sequence are masked, and the model is architected to learn to predict the deliberately left out (masked) words. We also experimented with language models that were pre-trained on psychology and healthcare-relevant texts (e.g. PsychBERT,[77] ClinicalBERT,[78] BioBERT.[79] However, we observed poor diagnosis classification performance after fine-tuning on our autism reports dataset. This is likely because these domain-specific language models were pre-trained on much smaller datasets.

On the other hand, long context models such as Longformer[24] with a context window of 4096 tokens did yield some promising report classification results (see Figure S1). However, the implementation of their attention mechanisms, such as sliding window attention in the case of Longformer, pose a challenge for effective interpretability in our present setting. In particular, the ability to produce a singular dense report embedding, while effective for classification, does not allow for the level of granularity required for effective interpretability of individual semantic elements. Once we zoom in on the sentence level, our unit of interpretability, enabled by our bespoke interpretability pipeline, RoBERTa's context window of 512 tokens is more than sufficient, achieving classification performance on par or exceeding these long context models.

Language model architecture enhancements

While pre-trained language models afford advanced natural language capabilities and thus enable ground-breaking text classification performance, their complex inner workings elude simple interpretation. In their unmodified state, it is challenging to understand

the motivating factors that would lead an out-of-the-box language model to classify a multi-page report as resulting in a diagnosis of autism. Aiming to address some interpretational shortcomings of widespread deep language models,[80–82] we decided to split each report into single sentence segments, as we described previously. By forcing the language model to focus its information processing capabilities on compact human interpretable units, such as sentences, we build inherent interpretability into the modeling objective from the beginning. These single sentence segments were fed into our pre-trained language model (i.e., keeping the previously obtained set of model parameters fixed) in order to extract as much semantic content as possible. Therefore, the internal representation of our pre-trained language model instantiates an autism-relevant sentence embedding.

From there, as an extension of this language model framework, we introduced a trainable single-head attention[83–85] neural network layer, which has the potential to selectively pinpoint and up-weight certain sentences contained within a report. The purpose of this layer is to refocus the representational capacity of the language model on the report-level, while retaining the ability to assess the influence of individual sentences on the autism diagnosis classification. Specifically, the attention mechanism projects the input matrix of autism-relevant sentence embeddings (from the fine-tuned language model) $X \in \mathbb{R}^{N_T \times d_{\text{emb}}}$ into key (K), query (Q), and value (V) subspaces using trainable weight matrices ($W^K, W^Q, W^V \in \mathbb{R}^{d_{\text{emb}} \times d_{\text{emb}}}$), where d_{emb} denotes the number of dimensions in the embedding space (768 in our case) and N_T refers to the number of tokens/words in a sentence. We then calculate the "attention budget" that the transformer layer affords to each individual sentence by all the other sentences in a report by multiplying the key and query vectors together (dot product multiplication). This produces a square attention score matrix $A \in \mathbb{R}^{N_S \times N_S}$, where N_S is the number of sentences in a report and the indices i, j correspond to the attention from sentence i on sentence j, in a given report, with the constraints:

$$A_{ij} \geq 0$$

$$\sum_{j=1}^{N_S} A_{ij} = 1$$

The value vectors are then multiplied by their corresponding attention scores from this attention matrix, which act as importance weights. We finally take the mean of these weighted sentence/value vectors to obtain a single embedding vector representation of each report document. In essence, the key, query, and value vectors serve as transformations of the input sentences into separate input-variant latent spaces, allowing us to automatically learn, and preferentially boost, meaningful relationships between the sentences in a report.

$$K = XW^K, K \in \mathbb{R}^{N_S \times d_{\text{emb}}}$$

$$Q = XW^Q, Q \in \mathbb{R}^{N_S \times d_{\text{emb}}}$$

$$V = XW^V, V \in \mathbb{R}^{N_S \times d_{\text{emb}}}$$

$$A = \text{Attention} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_{\text{emb}}}}\right)$$

$$X_{\text{output}} = AV, X_{\text{output}} \in \mathbb{R}^{N_S \times d_{\text{emb}}}$$

$$x_{\text{report}} = \frac{1}{N_S} \sum_{i=1}^{N} X_{\text{output}_i}$$

The benefits of this specialized language model architecture were two-fold. First, we directly ensure that the embedding representation of each report is composed of a diagnosis-contingent semantic mixture of sentences, from a report at hand, proportional to the importance of those sentences in terms of carrying information useful for detecting autism cases. Second, since we use a single attention head to weight the importance of each sentence in a report, we can directly interpret the unique attention matrix produced by the attention layer for each report. Sentences that were flagged as most highly attended according to this matrix were, by construction of our language model architecture, most important for the classification of text from reports with the autism diagnosis. These insights were derived on a per report basis, such that we could identify the most salient sentences in the context of a given report. In sum, we have segmented each report into individual sentences and used a pre-trained language model to extract the

semantic content from those sentences for the purpose of autism detection. We subsequently brought to bear a single-head attention layer to select and preferentially up-vote specific sentences that most contribute to the diagnosis classification.

Fine-tuning framework to spawn an autism-aware embedding representation

Capitalizing on our bespoke architecture, we next fine-tuned the pre-trained language model and our single-head attention layer on our corpus of 4272 reports on cases with suspected autism. The target modeling outcome of this fine-tuning was to classify the ultimate diagnosis of the patient from a single report alone. To accomplish this distinction between clinically confirmed versus suspected but initially ruled-out autism, we passed the attention-weighted report embeddings through a final fully connected layer followed by a sigmoid to issue an autism probability. This probability was then compared with the final diagnosis using a binary cross-entropy loss. This loss function comprised our training objective. It is important to reiterate that these reports do not contain any explicit or hinted judgment as to the eventual diagnosis, as all information regarding diagnoses and standardized scores were manually removed during pre-processing and the anonymization process for each report. In terms of hyperparameter choices, we fine-tuned on our training set for a total of 2 epochs, using a batch size of 8 reports at a time, and a learning rate of 1×10^{-5} with the AdamW optimizer.[86,87] Hyperparameters were chosen via grid search on a small held out validation set (5% of reports). Each sentence was truncated or padded to a length of 64 tokens, with padding tokens representing irrelevant entries. Further, the number of sentences per report was also truncated or padded to 64, using the same padding scheme. Embeddings (corresponding to words or sentences) were given a representational dimensionality of 768 dimensions.[88]

In order to allow our language model pipeline to act on sentences as the unit of inference and interpretation, we averaged the word embeddings across embedding dimensions from the output of our fine-tuned language model to produce sentence embeddings of identical dimensionality (768 dimensions in all cases). Then, once the language model word embeddings had been abstracted to sentence embeddings, our single-head attention module could act on the level of sentences to identify the most autism relevant sentences in a given report. Finally, sentences within each report document were combined into overall report embeddings, using a weighted average where the weights were ascertained by the single-head attention module.

Our model was fine-tuned in an end-to-end fashion: the parameter weights of the pre-trained language model, our single-head attention layer, and the final classification layer were all trained simultaneously. To ensure the validity of our accuracy results, we performed a rigorous 5-fold cross-validation procedure. Within each cross-validation fold, 80% of the reports were randomly chosen to comprise our training set for fine-tuning, while the remaining 20% of reports were set aside as our test set to evaluate the performance of our model on new, unseen reports. It is important to note that we grouped multiple reports from the same patient into *either* the training *or* the test sets, exclusively, to ensure independence between the training and test sets. That is, a given subject's information only ever appeared either exclusively in the training set or exclusively in the test set. A diagnosis classification accuracy for each cross-validation fold was calculated using the held-out test set at the end of the fine-tuning process; averaging across these 5 fold-wise accuracies yielded the prediction performance that is expected for new, unseen reports from the same underlying distribution. Baseline models included traditional bag of words (BOW) in which word counts for each word in a report were used as input features for a linear (naïve Bayes) or non-linear (random forest) classifier, and a Doc2Vec[23] model in which reports are projected into report embeddings using a simple neural network. Hyperparameters for these baseline models were also optimized via grid search and average accuracies were again computed with a rigorous 5-fold cross-validation procedure, in the same fashion as our language model validations. Larger transformer models such as Llama 3.1 8B[26] and Gemma 7B[27] were fine-tuned using low-rank adaptation (LoRA)[89] with a rank of 8 and an alpha parameter of 8. Zero shot prediction using these larger models incorporated a trained linear classifier (logistic regression) on top of the raw model embeddings; the parameters of the models themselves were not adjusted in this setting.

Using the GPU cluster of Mila Quebec AI Institute, we had the necessary compute resources to perform a full fine-tuning of our RoBERTa language model. However, LoRA would be an appropriate choice for fine-tuning much larger language models or for those without access to sufficient compute resources to perform full fine-tuning on RoBERTa. Since LoRA is a low-rank approximation of full fine-tuning, full fine-tuning as performed here represents an upper bound on the performance that could be attained for a given language model.[89,90] Regardless, fine-tuning of some kind is necessary to reach useful diagnostic classification performance. Simply attempting to classify the pre-trained embeddings generated from powerful language models such as Gemma 7B or Llama 3.1 8B without fine-tuning using a linear classifier (zero-shot learning) yielded only slightly above-chance classification performance on our corpus of reports (see Figure S1). Therefore, we concluded that fine-tuning appears to be necessary for the construction of an embedding space that allows for the direct comparison of sentences on the basis of their autism-relevance. Generic language model embedding spaces are demonstrably unsuitable for this task.

QUANTIFICATION AND STATISTICAL ANALYSIS

Tactics tailored for model interpretability

After successfully refining our language model on the corpus of autism reports, we wanted to isolate and interpret the driving elements of autism diagnosis, as verbalized by clinicians and as extracted by our model. First, we inputted our entire corpus of reports using our fine-tuned language model to generate sentence embeddings and sentence-instantiated attention matrices. Based on these sets of attention scores A, we get a sense of how the language model appraises the information content in articulated statements that retrace the clinical thought process — a necessary step toward providing a quantitative answer as to what really matters in

distinguishing a child with autism. With these embeddings for every sentence in our corpus, we next wanted to investigate and visualize the sentence embedding instances residing in the language model space. Specifically, we analyzed the 768-dimensional sentence embedding representations from the last hidden state (LHS) — corresponding to the 12th layer — of the pre-trained language model. We performed a PCA decomposition on these LHS embeddings to reduce the dimensionality to two dimensions for simple visualization.

Next, we sought to shed light onto the inner workings of the language model itself; specifically, we wished to bring to the surface the flow of information through the various layers, and reveal at what consecutive processing stage the model begins to extract meaningful semantic information useful for the autism diagnosis classification task. To accomplish this goal, we extracted the sentence embedding representations from each neural network layer (starting at the first layer, all the way to the 12th and final layer), and computed their mean per report, across the 768 embedding dimensions, to generate an overall report embedding. We then used these layer-wise report embeddings to train linear classifiers (logistic regression models, one model per layer) to again predict the diagnosis. Similarly to our language model fine-tuning evaluation, we implemented a rigorous 5-fold cross-validation scheme to evaluate these logistic regression models on unseen report embeddings. To evaluate the layer-wise model performance, we computed an out-of-sample AUC score by averaging the AUC across folds for each layer individually.

Subsequently, we endeavored to analyze the semantic content of the reports themselves. We began by identifying the most ''important'' sentence per report as pinpointed by our single-head attention layer. The rule for identifying the most important sentence was straightforward: whichever sentence had the highest overall attention score across all other sentences in a report, as calculated by summing the rows of attention matrix A, would be deemed the most ''important''.

$$\max_{j} \sum_{i=1}^{N_S} A_{i,j}$$

where i and j denote the rows and columns of the attention matrix A (size $N_S \times N_S$, cf. above notation), respectively. We next wanted to identify broad features in the most important sentences in each report that differ between patients that were diagnosed with autism and those sentences that were less discriminatory. To examine these differences, we counted the number of times a given word appeared in the top attended sentence in all the reports. Next, for each word we divided the total number of occurrences of that word in the autism-associated report top sentences by the total number of occurrences of that same word in the top sentences from reports not associated with an autism diagnosis. These totals were calculated across our entire corpus of reports. This approach yielded a relative frequency differential for the usage of certain words in the most autism-relevant sentences across reports, of direct interpretational value.

Evaluation against external diagnostic criteria

For our final set of analysis steps, we wanted to assess the utility of external diagnostic criteria in the context of our language model-parsed clinical intuition, as a litmus test. Further, we wanted to perform an explicit comparison of these external criteria to our language model-identified semantic elements contained within our corpus of reports. The rationale for this analysis was that the decision-making process of language models can be inherently difficult to interpret, just by itself. To overcome some shortcomings of currently existing language model frameworks,[20] we anticipated that by bringing well-established external knowledge to the table and juxtaposing it with the formed autism-aware semantic embedding space spanned by our language model, we could potentially throw light onto the inner workings of our language model as well as, enabled by this explainability approach, empirically revisit these standard clinical guidelines.

In particular, since our language model is, by its nature, able to process any natural language input, we were able to generate sentence embeddings from text descriptions of each of the DSM-5 autism criteria. Crucially, we could then evaluate the established criteria explanation embeddings by comparing them to the embeddings of our previously identified diagnosis-critical sentences from our reports. Hence, this explainability strategy enabled us to bring together the embeddings of the widely accepted DSM-5 catalog and those of our present report sentences, to make sense of what our language model did in fact learn during the model fine-tuning process. To carry out this comparison against externally established sources of knowledge, we turned to the cosine similarity metric. This metric has been used extensively in a variety of NLP contexts to evaluate the semantic similarity between any two embeddings.[29,30] Conceptually, the rationale behind this metric is that vectors pointing in similar directions in the high-dimensional semantic embedding space are taken to carry similar meanings. Practically, a cosine similarity of -1 indicates opposite semantic information, 0 indicates no relation in meanings, and +1 indicates identical semantic content for a given pair of sentence embeddings. Thus, this metric served as an easily interpretable tool for assessing the similarity of specific sentences, in the context of the autism-aware embedding space of our language model. The cosine similarity of any two semantic embedding vectors b and c is found by calculating the cosine of the angle between them:

$$cos(\theta) = \frac{b \cdot c}{||b||||c||} = \frac{\sum_{i=1}^{d} b_i c_i}{\sqrt{\sum_{i=1}^{d} b_i^2}\sqrt{\sum_{i=1}^{d} c_i^2}}$$

where *d* denotes the dimensionality of the semantic embedding vectors, and *i* serves as an index for each dimension. Precisely, we calculated cosine similarities between all seven of the DSM-5 autism criteria, and the most diagnosis-relevant sentence in each report. In so doing, we associated each report in our corpus with seven distinct cosine similarity scores, despite the fact that the DSM-5 criteria were never administered in our participant cohort.

While looking at the distributions of the raw values of the computed similarity scores between non-diagnosed and diagnosed autism groups of children was informative, we wanted to directly validate that these similarity scores were truly meaningful for autism diagnosis. To achieve this goal, we fitted an LDA model to once again predict the diagnosis for each report. This LDA model was based on an implementation from the scikit-learn Python package. LDA is a model that aims to find a linear combination of input features that best separates the groups while maximizing the between-group variance and minimizing the within-group variance (in our case, a diagnosis of autism or a non-confirmed diagnosis).[91] As input features to this LDA model, we used the input feature vector of seven cosine similarities (for each report) corresponding to the seven DSM-5 criteria. In modeling the data, LDA assumes that the group-conditional probability distributions are two normal distributions with differing means μ_1, μ_2 and covariance Σ shared between the two groups. The decision boundary is thus the solution (y) to the following linear equation:

$$log\ \pi_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + y^T\Sigma^{-1}\mu_1 = log\ \pi_2 - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + y^T\Sigma^{-1}\mu_2$$

where μ_1 and μ_2 are the group-wise means for each of our 7 cosine similarities, Σ is the covariance matrix shared across internal class representations, and π is a constant indicating the prior proportion of the labels for each respective group in our dataset. By means of this LDA model, we wanted to concretely evaluate how the set of the 7 DSM-5 criteria, in the context of our autism-aware internal language model representation, contribute to achieving an accurate autism classification in children. We assessed this LDA model trained on the DSM-5 criteria cosine similarities using a rigorous 5-fold cross-validation framework, in an identical format to our previous cross-validation schemes, and evaluated the model performance (AUC) in telling the two groups apart for each fold on unseen independent reports. To determine the direction of association for each of the DSM-5 criteria similarities, we correlated the raw cosine similarities of each criterion with the LDA linear combination of all of these inputs, that is the LDA "scores" for each observation. Since this linear combination is definitionally the most discriminative transformation of the data under the LDA model, the directionality and strength of these correlations revealed which DSM-5 criteria in our language model space are most useful for predicting a diagnosis of autism, conditioning on the previously derived semantic embedding space. These ensuing results served to confirm the prior analysis of the raw group-wise cosine similarities for each of the DSM-5 criteria.

As a negative test, to be certain that our most attended sentences per report were indeed uniquely useful for the diagnosis of autism and the comparison to external criteria, we reiterated the entire LDA procedure, but this time selected a random sentence from each report instead of the top attended sentence and calculated the cosine similarities between these random sentences and the DSM-5 criteria. We then re-trained and re-evaluated our LDA model on these new cosine similarities. This negative control confirmed that our attention-selected important sentences were in fact particularly useful for predicting a diagnosis of autism and as a comparison to external diagnostic criteria, in this sensitivity analysis. Taken together, these out-of-sample validation investigations served to confirm the fundamental soundness of the key autism-critical sentences identified by our explainable language model architecture, in addition to confirming the generalized trend of the directionality of the established DSM-5 criteria with regards to autism diagnosis. Through this methodology, we were able to successfully concretize abstract concepts in a tangible and fully quantitative fashion, to empower rigorous comparison and evaluation of those concepts.

Supplemental figures

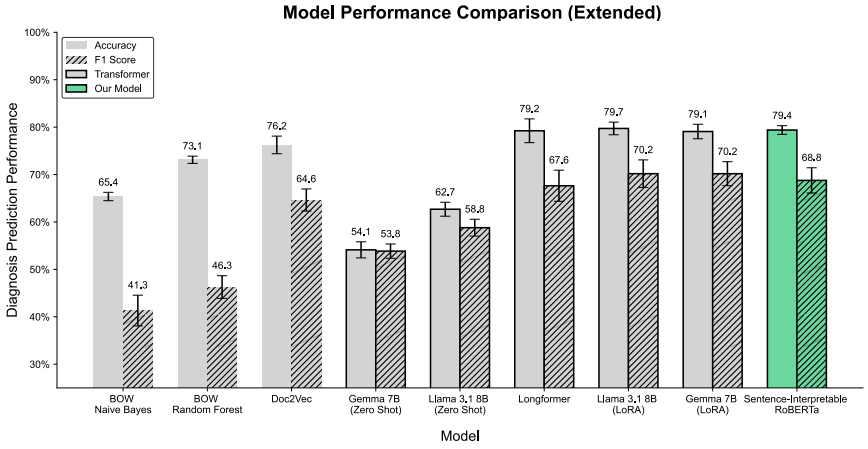


Figure S1. Transformer language models offer performance gains on autism diagnosis classification after fine-tuning on healthcare professional reports, related to Figure 2

Out-of-sample benchmarking of predicting autism diagnosis from unstructured healthcare professional reports. Bar height indicates the average prediction performance (accuracy or F1 score), while the whiskers indicate the variability (one standard deviation) of the metric after fitting the model independently in five different cross-validation folds. Solid-colored bars correspond to classification accuracy, whereas bars with diagonal stripes correspond to F1 score, for the same model. Transformer-based language models have been outlined in black, and our custom sentence interpretable language model is shaded green. As in Figure 2A of the main text, legacy NLP approaches (bag of words [BOW], Doc2Vec), have been compared with various leading transformer model architectures. These modern language models have been evaluated in a zero-shot setting (raw embeddings followed by a fitted linear classifier), as well as following fine-tuning on our corpus of text reports (using low-rank adaptation [LoRA]).

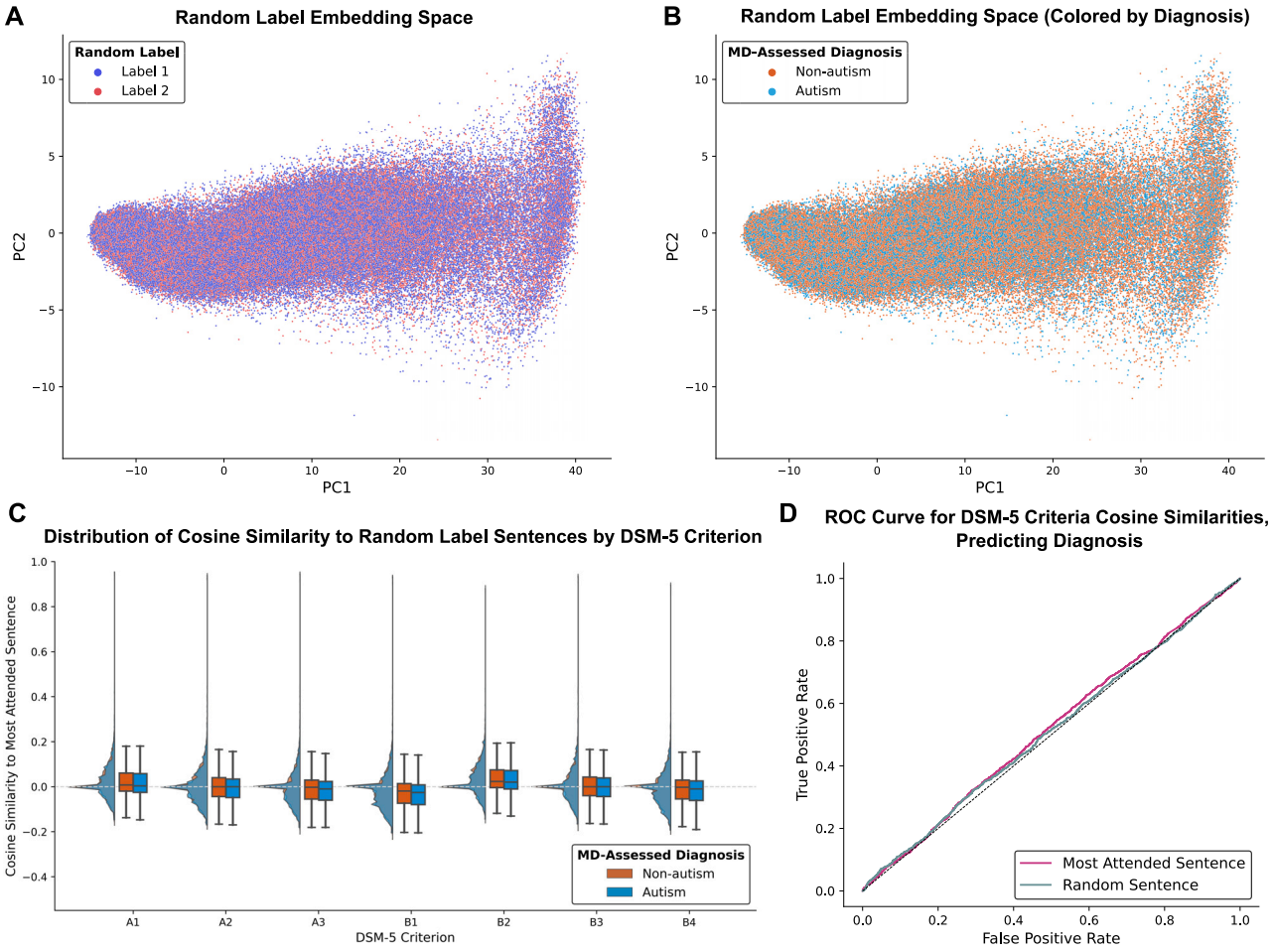


Figure S2. Predicting random outcomes in our original transformer model yields random results, related to Figure 5

(A) After generating sentence-level embeddings from our random label fine-tuned language model for every report, a PCA decomposition of these embeddings does not show obvious structure. Each point in the PCA plot corresponds to a single sentence and is colored according to the random "diagnosis" label.

(B) Identical PCA plot of sentence-level embeddings as shown in (B), now with each sentence embedding colored by the actual autism diagnosis (cf. Figure 2C in the main text). No further structure is revealed after performing this labeling.

(C) After passing each DSM-5 autism criterion description (A1–B4), through our random label fine-tuned model and obtaining a sentence-level embedding representation for each criterion, we obtain a distribution of cosine similarities between the embedding of each criterion and the most attended sentence for each of our reports. This distribution is represented by a density plot as well as a boxplot corresponding to the interquartile range for the cosine similarities between the DSM-5 criteria and the most attended sentence in each report. These distributions are divided based on the final diagnosis for each report (autism versus non-autism).

(D) ROC curve showing the out-of-sample classification performance on unseen reports of the LDA model trained on the cosine similarities of each DSM-5 criterion and the most attended sentences in each report. These cosine similarities to the most attended sentences (purple line) perform just as poorly on this diagnosis classification task as cosine similarities to random sentences from each report (teal line). This reveals that these cosine similarities are not meaningfully relevant to autism diagnosis classification.

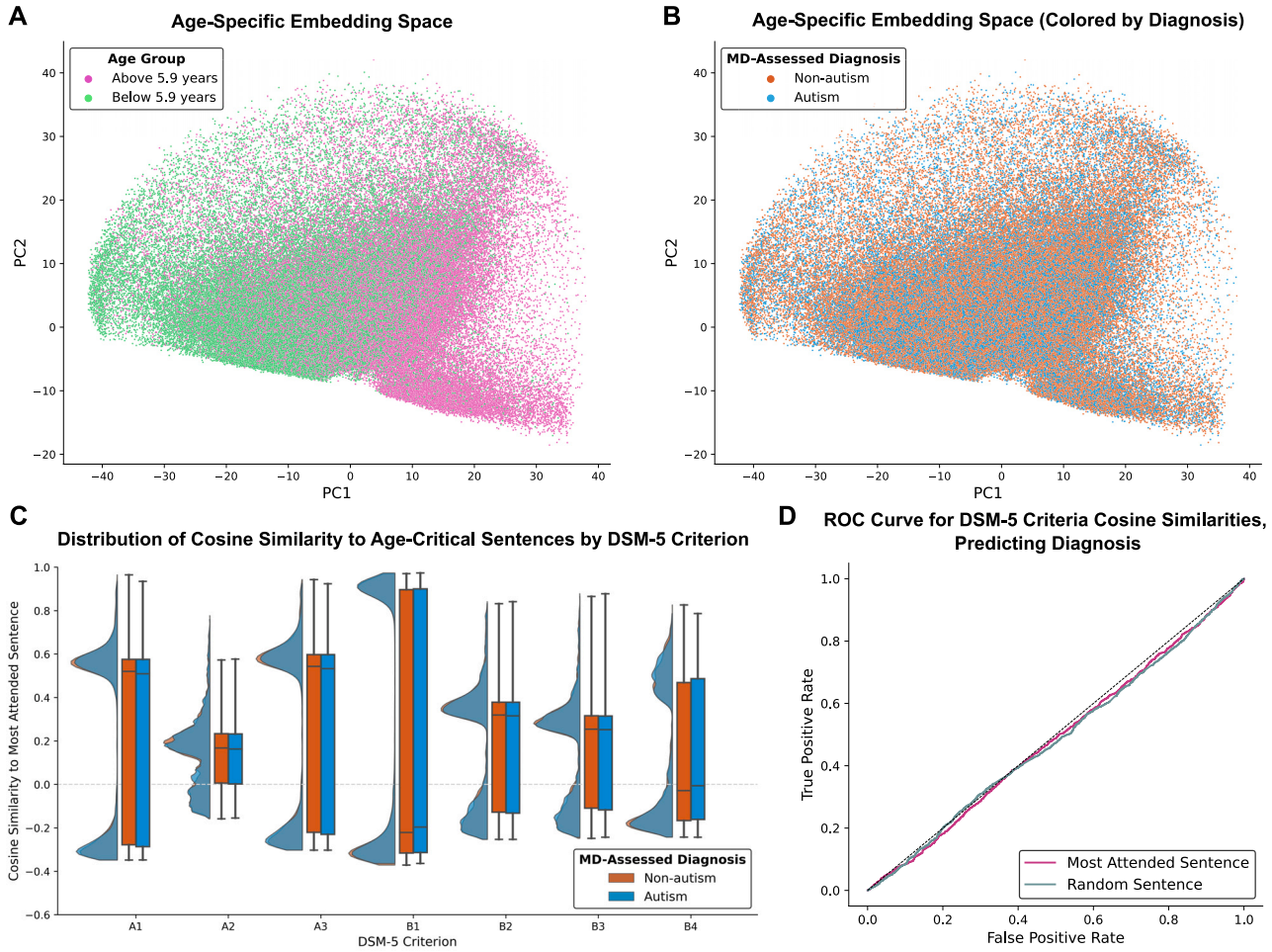


Figure S3. Predicting age differences in our original transformer model yields autism-irrelevant results, related to Figure 5

(A) After generating sentence-level embeddings from our age group fine-tuned language model for every report, a PCA decomposition of these embeddings shows some clear age-specific structure. Each point in the PCA plot corresponds to a single sentence and is colored according to the age group label.

(B) Identical PCA plot of sentence-level embeddings as shown in (B), now with each sentence embedding colored by autism diagnosis (cf. Figure 2C in the main text). While there may be age-specific structure in this embedding space, this generated embedding space does not appear to be autism-relevant.

(C) After passing each DSM-5 autism criterion description (A1–B4), through our age group sensitive model and obtaining a sentence-level embedding representation for each criterion, we obtain a distribution of cosine similarities between the embedding of each criterion and the most attended and age-critical sentence for each of our reports. This distribution is represented by a density plot as well as a boxplot corresponding to the interquartile range for the cosine similarities between the DSM-5 criteria and the most attended sentence in each report. These distributions are divided based on the final diagnosis for each report (autism versus non-autism).

(D) ROC curve showing the out-of-sample classification performance on unseen reports of the LDA model trained on the cosine similarities of each DSM-5 criterion and the most attended sentences in each report. These cosine similarities to the most attended sentences (purple line) perform just as poorly on this diagnosis classification task as cosine similarities to random sentences from each report (teal line). In this control analysis, the cosine similarities are hence not meaningfully relevant to autism diagnosis classification.