Mode-Aware Phenotype Profiling from Korean Clinical Reports: An LLM-Derived Two-Layer Fingerprint for Autism Characterization

Yeojin Ryu [1 2 3] **Miae Oh** [4] **Joon-Yong An** [1 2 3 5 6]

Abstract

Autism spectrum disorder (ASD) is phenotypically heterogeneous. Clinical diagnostic reports provide a richer and more clinically proximal basis for heterogeneity analysis than questionnaire totals. We analyzed 309 Korean clinical reports with a sentence-attention KLUE-RoBERTa model and an LLM silver-labeling workflow that mapped high-attention clinical sentences into 18 DSM-5-grounded phenotype domains. Report-level domain frequency vectors agreed closely with expert-gold profiles. To test for discrete subtypes, we fit Gaussian mixture models on isometric log-ratio coordinates and benchmarked the resulting clusters against a permutation null with multi-seed stability checks. Four *weak density modes* emerged consistently across random seeds but were not well separated enough for hard subtype assignment, comprising a dual prototype core, a broad diffuse mode, and a physiological-adaptive directional mode. Mahalanobis tail tests then located individual idiosyncrasy in the residual PC subspace rather than in the leading mode-carrying dimensions. We summarize the phenotype space with a two-layer fingerprint, mode membership plus within-mode tail deviation per domain, as a descriptive alternative to hard subtype labels, with the individual-deviation layer remaining exploratory.

[1]Department of Integrated Biomedical and Life Science, Korea University, Seoul, Republic of Korea [2]L-HOPE Program for Community-Based Total Learning Health Systems, Korea University, Seoul, Republic of Korea [3]National Research Laboratory for Convergence Degradation Biology, Korea University, Seoul, Republic of Korea [4]Department of Psychiatry, Kyung Hee University Hospital, Seoul, Republic of Korea [5]School of Biosystem and Biomedical Science, College of Health Science, Korea University, Seoul, Republic of Korea [6]Interdisciplinary Major Program in Targeted Degradationbased Innovative Therapeutics, Korea University, Seoul, Republic of Korea. Correspondence to: Joon-Yong An <joonan30@korea.ac.kr>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

Autism spectrum disorder (ASD) is clinically defined by social-communication differences and restricted or repetitive behaviors, but two people with the same diagnosis can differ sharply in cognition, sensory processing, adaptive function, co-occurring psychiatric symptoms, and developmental trajectories (Lord et al., 2022; Lombardo et al., 2019). Data-driven attempts to organize this heterogeneity have produced cluster solutions whose number and composition shift across cohorts and modalities (Narita et al., 2020; Lefort-Besnard et al., 2020; Zhao et al., 2024; Litman et al., 2025), and comparative latent-structure analyses argue that ASD phenotypes look more multi-dimensional than discrete (Kim et al., 2019; Frazier et al., 2023). Whether ASD heterogeneity is structurally categorical at all is still debated (Mottron & Bzdok, 2020; Lin et al., 2026). Normative models offer one alternative, expressing each individual as a deviation from a continuous reference rather than a cluster label (Marquand et al., 2016; Zabihi et al., 2019; 2020), and connectivity studies have shown that residual-level idiosyncrasy can dominate group-mean differences in autism imaging (Benkarim et al., 2021).

Clinical diagnostic reports record the actual observations clinicians weigh in making the ASD diagnosis. Where scale instruments compress diverse behavioral observations into a few numerical dimensions, narrative reports preserve the granularity of those observations even after structured extraction into domain vectors. Sentence-attention language models can recover diagnosis-relevant statements from such reports (Stanley et al., 2025), but the structure of the resulting phenotype space, and how to summarize that structure for an individual patient, are still open questions.

We work with Korean clinical reports and an LLM-derived 18-domain phenotype representation. We look at the phenotype space along two axes, density structure from mixture modeling and tail behavior from residual dimensions, and propose a mode-aware, within-mode deviation summary that refines the prototype-plus-idiosyncrasy hypothesis (Lin et al., 2026).

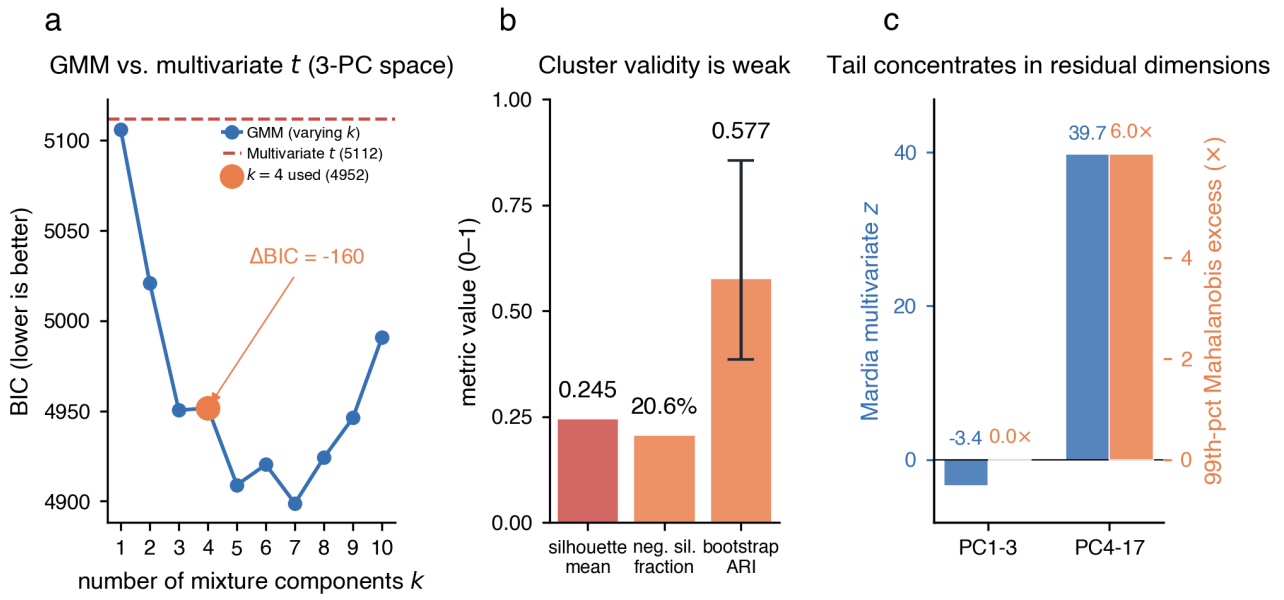

Figure 1. Phenotype-space diagnostics. (**a**) A four-component GMM improves BIC over a multivariate t model, but (**b**) low silhouette, frequent negative assignments, and moderate bootstrap ARI (with 95% CI error bar) show the modes are non-random yet not well-separated. (**c**) Tail structure concentrates in the residual subspace PC4–17 rather than the leading PC1–3.

2. Methods

Data and clinical language model. We analyzed 309 de-identified Korean clinical reports, 248 ASD and 61 non-ASD, from a single tertiary site. Reports came from two documentation streams, autism diagnostic reports (A-type, $n = 233$) and psychological assessment reports (P-type, $n = 76$). KLUE-RoBERTa (Park et al., 2021) was fine-tuned for binary ASD vs. non-ASD prediction with stratified five-fold CV. Following Stanley et al. (2025), each report was encoded as a sequence of sentence embedding vectors. A single-head sentence-level self-attention module transformed these vectors before valid-sentence mean pooling produced the report representation. The resulting sentence-by-sentence attention matrix was summarized into per-sentence importance scores, which were used to select high-attention sentences for phenotype labeling. High-attention sentences for each report were selected from the out-of-fold model in the five-fold procedure, so no report's phenotype vector was generated using attention scores from a model trained on that report.

Eighteen-domain phenotype profiles. High-attention sentences were labeled into 18 phenotype domains grounded in DSM-5 ASD criteria and pediatric psychiatric categories with Llama-3.1-8B run locally for de-identification compliance. Each sentence was forced into exactly one domain, with an explicit other-general option when no domain clearly applied, returned in a fixed JSON schema with a confidence value and a short rationale. We drew five independent votes per sentence at temperature 0.7 and seeds 42 to 46, discarded votes below a confidence of 0.5, and adopted a domain only when it won a majority of at least three of the five votes, assigning sentences with no majority to other-general. Per-report 18-domain frequency vectors are the analysis unit. We validated this representation against expert-gold annotations on 15 hand-labeled reports at the report-vector level, and report the five-vote self-consistency agreement in Appendix B.

Compositional phenotype-space analysis. Domain-frequency vectors from the 248 ASD reports lie on the simplex, where Euclidean operations are unsound, so we mapped them to isometric log-ratio (ILR) coordinates on a Helmert basis (Aitchison, 1986; Egozcue et al., 2003) before reducing to three PCs that explained 44.0% of within-ASD variance. We compared Gaussian mixture models with $k = 1$ to 10 against a multivariate t via BIC. Cluster geometry was assessed with silhouette, bootstrap ARI, and 50-seed pairwise ARI, and contrasted with a 1,000-permutation null. Tail behavior was quantified with Mahalanobis distance and Mardia multivariate kurtosis. Domain signals within each mode were classified by two size-matched permutation tests, comparing each mode's high-Mahalanobis members with random subsets drawn either from the full ASD cohort or from the same mode, with BH-FDR correction at $q \leq 0.10$. The combined view of weak modes and individual deviations follows the normative-model approach in clinical neuroscience (Marquand et al., 2016).

Table 1. Density modes in the LLM-derived phenotype space. Counts are ASD reports per mode. R = real signal, I = cluster identity, T = tail-specific. The clinical reading is a descriptive summary of domain emphasis for clinician interpretation, not a validated subtype.

Mode	n	Geometry	Signals (R/I/T) *	Clinical reading †
M1	70	central	T: insistence on sameness	Prototype core; sameness-driven tail
M2	23	central	none tested ‡	Diffuse prototype; no single signal
M3	97	diffuse	R: externalizing, internalizing	Behavioral co-occurrence (exploratory)
M4	58	directional	R: physiological function	Physiological–adaptive; P-type–linked

*BH-FDR $q \leq 0.10$ on two permutation tests (full ASD background and within-mode tail strata). †Descriptive, not a validated category; see §5. ‡n_{tail} below threshold.

3. Results

3.1. Diagnostic Model Performance

The fine-tuned classifier reached five-fold CV mean AUROC 0.910 (95% CI 0.838–0.984), with per-fold AUROC 0.749–0.987 and a frozen-layer probe peaking at Layer 9 (AUC 0.894); see Appendix A. High-attention sentences from this classifier formed the per-report pool used for 18-domain silver labeling and all downstream phenotype analyses, so the classifier functions as a sentence selector rather than a diagnostic deployment claim.

3.2. Modes are Significant but Boundaries are Weak

A four-component GMM fit the leading phenotype space better than a multivariate t model. GMM BIC was 4951.7, t BIC was 5112.1, so ΔBIC was -160 in favor of GMM (Figure 1). Read on its own, this would suggest four latent groups.

Geometry and stability gave a different picture. Average silhouette was only 0.245, with 20.6% of samples receiving negative assignments, and bootstrap ARI averaged 0.577 with 95% CI 0.386 to 0.856. The multivariate t fit converged to $\nu = 500$, effectively Gaussian, so the leading PC1–3 space is not well described as a single heavy-tailed cloud.

To separate "weak modes" from "no signal," we compared the observed multi-seed pairwise ARI from 50 seeds (mean 0.907, 95% CI 0.49 to 1.00) against a 1,000-permutation null on cluster labels (mean ≈ 0, Appendix C). Multi-seed agreement was far above this label-permutation baseline, indicating that the GMM solution is algorithmically stable relative to random labelings. We therefore interpret the components as weak recurrent density modes, not as evidence for discrete subtypes. Per-mode persistence across 50 seeds ranges 1.00 (M4) to 0.95 (M3). BIC was shallow across $k = 3$ to 7, varying by about 1%, so model fit alone did not single out a value of k. We fixed $k = 4$ as a descriptor, weighing fit together with cross-seed stability (Appendix C).

3.3. Idiosyncrasy Lives in Residual Dimensions

Individual specificity was strongest outside the principal modes (Figure 1c). PC1 to PC3 explained 44.0% of within-ASD variance and were mildly platykurtic (Mardia $z = -3.41$), indicating that the leading mode-carrying dimensions are if anything more concentrated than Gaussian. The residual PC4 to PC17 subspace explained 56.0% and was sharply leptokurtic (Mardia $z = 39.69$, 99th-percentile Mahalanobis excess 6.05% vs. expected 1%). Coarse modes thus capture recurrent clinical emphases, while patient-specific deviations concentrate in residual dimensions.

Controlling for report type leaves this residual tail intact, with the PC4 to PC17 Mardia z at 37.6 after control and 39.7 before, and still favors a mixture over a single multivariate t. The four modes weaken further under the same control, which fits their reading as weak density modes rather than discrete subtypes (Appendix F).

3.4. Four Density Modes with Sparse, Mode-Dependent Signals

As a soft coordinate system, the four modes were clinically informative (Table 1). Two central modes, M1 and M2, anchored the prototype core, holding 93 of the 248 ASD reports together. M3 was a diffuse central cloud, the largest mode but with the most dispersed members. M4 captured a physiological-adaptive directional profile reaching into the tail.

The four-quadrant signal analysis (Figure 2) showed that each mode carried a small number of distinguishing domains rather than a clean profile separation. ASD-core domains carried most of the validated mode-defining signals, while peripheral psychiatric and format domains contributed less stably and showed substantially lower fidelity to gold annotations (Appendix B). Because M3's two validated signals, externalizing and internalizing, both fall in this less reliable group, its mode-defining profile should be read as an exploratory descriptor rather than a set of validated clinical markers. The M2 prototype mode ($n = 23$) was below the size threshold for testing, and M1 carried only a tail-specific insistence-on-sameness signal. If clusters were natural subtypes, broad domain profiles would cleanly separate the modes. Instead, each mode contains a handful of interpretable signals embedded in a shared ASD background.

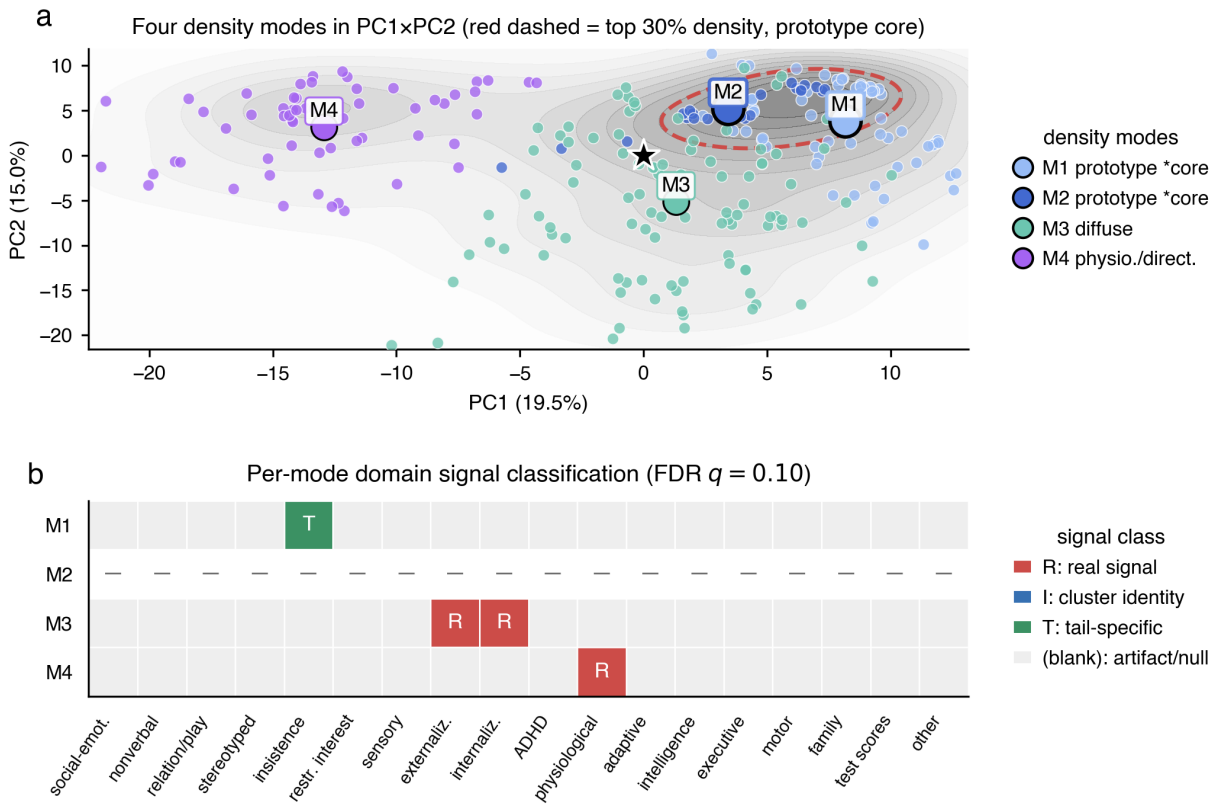

Figure 2. Four weak density modes and their sparse domain signals. (**a**) ILR PC1×PC2 density landscape (ASD $n = 248$, KDE contours in grey): each report is colored by its $k = 4$ GMM mode assignment, large outlined markers indicate mode centroids, and the red dashed contour encloses the top 30% density region. Modes M1 and M2 anchor the prototype core within this region ($n = 93$); the black star marks the ASD centroid. (**b**) Per-mode 18-domain signal classification after Benjamini–Hochberg FDR correction ($q \leq 0.10$). Domain names are abbreviated, with full names and validation in Appendix B. R = real signal, I = cluster identity, T = tail-specific, blank = artifact/null.

4. A Two-Layer Fingerprint

In the full 17-dimensional ILR space, hard mode labels account for 24.7% of phenotype variance, with the remaining 75.3% residual within modes (label-permutation null 1.2%); in the leading 3-PC space where the modes are defined they account for 54.2% (Appendix D). We summarize the phenotype space with a two-layer fingerprint, whose second layer describes this within-mode residual, as an alternative to unstable subtype labels. Layer 1 assigns each patient a soft density-mode membership (GMM posterior over $k = 4$). Layer 2 describes domain-level deviations relative to that mode, classified as real, cluster-identity, or tail-specific by the same permutation tests used at the cohort level (FDR $q \leq 0.10$), and marks the domains on which a patient departs from others in the same mode rather than assigning an individual subtype label. For each patient, the within-mode z-score on each domain is computed relative to reports assigned to the same mode, and the fingerprint highlights a domain when the patient's $|z|$ is large and the permutation tests (Table 1) classify that domain as a mode-informative signal.

As an illustration, one anonymized M1-mode patient in our cohort has a GMM posterior of 0.95 for M1 and a Mahalanobis distance of 5.5 in ASD phenotype space (top-decile tail), with a within-M1 z-score of 3.0 on insistence on sameness, a domain that itself surfaces as a tail-specific signal in M1 (Table 1). The fingerprint reads "prototype mode, high confidence, with within-mode tail deviation on insistence on sameness," rather than collapsing to a categorical "M1 subtype."

5. Discussion

Our analyses describe a phenotype landscape that is structured but soft. Modes are stable relative to permutation null but their boundaries are not sharp enough to anchor patient-level labels (Kim et al., 2019; Frazier et al., 2023). Both observations together push toward a representation-level summary of patient phenotypes rather than a classification-level one (Mottron & Bzdok, 2020; Marquand et al., 2016).

Relative to large phenotypic mixture studies (Litman et al., 2025) and large genotype-to-phenotype analyses (Rolland et al., 2023), this text-derived setting on a real-world clinical-report cohort surfaces weak density modes rather than well-separated classes. Relative to the prototype-plus-idiosyncrasy hypothesis (Lin et al., 2026) and normative-deviation parsing of autism (Zabihi et al., 2019; 2020), the data point to a dual prototype core plus sparse mode-defining signals, with idiosyncrasy concentrated in residual dimensions. The within-ASD representation spans about 12 to 13 effective dimensions out of 17, and stays this high after report type is regressed out (Appendix D), so a one-dimensional severity score would discard most of the variation that the report vectors carry.

Clinical reading of the modes. Read descriptively rather than as subtypes, the four modes correspond to recognizable emphases in the reports (Table 1). M1 and M2 form a prototype core dominated by ASD-core domains, with M1 distinguished by a tail-specific insistence-on-sameness signal and M2 a diffuse prototype carrying no single mode-defining domain. M3, the largest and most dispersed mode, is characterized by co-occurring externalizing and internalizing emphases; because both domains are weakly validated against gold annotations (Appendix B), this reading is exploratory. M4 is a physiological–adaptive directional mode, but it is composed almost entirely of psychological-assessment (P-type) reports, so its physiological-function signal may reflect documentation style rather than patient-level phenotype and should not be read as a clinical subtype. These readings are descriptive summaries of domain emphasis intended for clinician interpretation, not validated clinical categories; external clinical-variable validation is left to future work.

Why subtyping fails to replicate. If the underlying phenotype landscape contains several modes with diffuse rather than discrete boundaries, any cohort-specific clustering will mostly recover how those weakly anchored modes happen to surface in that sample, not the population-level structure. Forcing a fixed value of k on top of this further compounds the instability. This explains why questionnaire- and connectivity-derived subtype solutions yield non-overlapping numbers and compositions across studies (Narita et al., 2020; Lefort-Besnard et al., 2020; Zhao et al., 2024). A representation-level fingerprint that reports mode posteriors and within-mode deviation sidesteps the problem without committing to a categorical label.

Limitations. The cohort is single-site, Korean-language, $N = 309$, and cross-sectional. Silver labels carry sentence-level annotation noise that aggregates out at the report-vector level used for downstream analysis (Appendix B), with stronger fidelity for ASD-core domains than for peripheral psychiatric domains. Mode-defining signals attributed to weakly-validated domains, in particular externalizing and internalizing in M3, should be regarded as preliminary. Per-domain silver validation rests on 15 gold-annotated reports, which limits precision for rare domains such as externalizing and physiological function, and a larger gold set together with a larger cohort is the main route to a more robust silver pipeline. The cohort mixes two documentation streams, autism diagnostic and psychological assessment reports, which carry distinct clinical content. Validation on independent cohorts and other report types remains future work. The within-mode deviation layer is exploratory at the individual level: the silver-vs-gold Mahalanobis ranking is weakly correlated (Appendix B), so per-patient deviations should be confirmed against independent clinical variables before clinical use.

Ethics Statement

This study followed the principles of the Declaration of Helsinki and was approved by the Institutional Review Board of Kyung Hee University Hospital (KHUH 2026-01-086). As a retrospective review of existing clinical records, it was granted a waiver of informed consent, and all reports were de-identified before analysis. The fingerprint is intended as a descriptive phenotype summary for research use only, not for individual diagnosis or treatment recommendation.

Acknowledgements

This research was supported by the National Research Foundation of Korea under Grant Number RS-2024-00439474 (J-YA).

References

Aitchison, J. *The Statistical Analysis of Compositional Data*. Chapman and Hall, London, 1986.

Benkarim, O., Paquola, C., Park, B.-y., Hong, S.-J., Royer, J., et al. Connectivity alterations in autism reflect functional idiosyncrasy. *Communications Biology*, 4:1078, 2021. doi: 10.1038/s42003-021-02572-6.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35:279–300, 2003. doi: 10.1023/A:1023818214614.

Frazier, T. W., Klingemier, E. W., Anderson, C. J., Gengoux, G. W., Youngstrom, E. A., and Hardan, A. Y. Categorical versus dimensional structure of autism spectrum disorder: a multi-method investigation. *JCPP Advances*, 3(3): e12142, 2023. doi: 10.1002/jcv2.12142.

Kim, H., Keifer, C. M., Rodriguez-Seijas, C., Eaton, N. R., Lerner, M. D., and Gadow, K. D. Quantifying the optimal structure of the autism phenotype: a comprehensive comparison of dimensional, categorical, and hybrid models. *Journal of the American Academy of Child and Adolescent Psychiatry*, 58(9):876–886, 2019. doi: 10.1016/j.jaac.2018.09.431.

Lefort-Besnard, J., Vogeley, K., Schilbach, L., Varoquaux, G., Thirion, B., Dumas, G., and Bzdok, D. Patterns of autism symptoms: hidden structure in the ADOS and ADI-R instruments. *Translational Psychiatry*, 10:257, 2020. doi: 10.1038/s41398-020-00946-8.

Lin, H.-Y., Breakspear, M., and Mottron, L. From heterogeneity to idiosyncrasy in the autistic brain. *Nature Mental Health*, 4:346–359, 2026. doi: 10.1038/s44220-026-00601-z.

Litman, A., Sauerwald, N., Green Snyder, L., Foss-Feig, J., Park, C. Y., Hao, Y., Dinstein, I., Theesfeld, C. L., and Troyanskaya, O. G. Decomposition of phenotypic heterogeneity in autism reveals underlying genetic programs. *Nature Genetics*, 57:1611–1619, 2025. doi: 10.1038/s41588-025-02224-z.

Lombardo, M. V., Lai, M.-C., and Baron-Cohen, S. Big data approaches to decomposing heterogeneity across the autism spectrum. *Molecular Psychiatry*, 24:1435–1450, 2019. doi: 10.1038/s41380-018-0321-0.

Lord, C., Charman, T., Havdahl, A., et al. The lancet commission on the future of care and clinical research in autism. *The Lancet*, 399(10321):271–334, 2022. doi: 10.1016/S0140-6736(21)01541-5.

Marquand, A. F., Rezek, I., Buitelaar, J., and Beckmann, C. F. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biological Psychiatry*, 80(7):552–561, 2016. doi: 10.1016/j.biopsych.2015.12.023.

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

Narita, A., Nagai, M., Mizuno, S., et al. Clustering by phenotype and genome-wide association study in autism. *Translational Psychiatry*, 10:290, 2020. doi: 10.1038/s41398-020-00951-x.

Park, S., Moon, J., Kim, S., et al. KLUE: Korean language understanding evaluation. In *Advances in Neural Information Processing Systems*, 2021.

Rolland, T., Cliquet, F., Anney, R. J. L., Moreau, C., Traut, N., et al. Phenotypic effects of genetic variants associated with autism. *Nature Medicine*, 29(7):1671–1680, 2023. doi: 10.1038/s41591-023-02408-2.

Stanley, J., Rabot, E., Reddy, S., Belilovsky, E., Mottron, L., and Bzdok, D. Large language models deconstruct the clinical intuition behind diagnosing autism. *Cell*, 188(8):2235–2248.e10, 2025. doi: 10.1016/j.cell.2025.02.025.

Zabihi, M., Oldehinkel, M., Wolfers, T., Frouin, V., Goyard, D., Loth, E., et al. Dissecting the heterogeneous cortical anatomy of autism spectrum disorder using normative models. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 4(6):567–578, 2019. doi: 10.1016/j.bpsc.2018.11.013.

Zabihi, M., Floris, D. L., Kia, S. M., Wolfers, T., Tillmann, J., et al. Fractionating autism based on neuroanatomical normative modeling. *Translational Psychiatry*, 10:384, 2020. doi: 10.1038/s41398-020-01057-0.

Zhao, W., Le, J., Liu, Q., et al. A clustering approach identifies an autism spectrum disorder subtype more responsive to chronic oxytocin treatment. *Translational Psychiatry*, 14:312, 2024. doi: 10.1038/s41398-024-03025-4.

A. Diagnostic Classifier Details

KLUE-RoBERTa-base (Park et al., 2021) was fine-tuned end-to-end with a single sentence-attention head over 153-sentence/128-token reports (lr 1×10^{-5}, AdamW, batch 1, dropouts 0.1, weight decay 0.01, up to 40 epochs with patience 10, seed 42). Stratified five-fold CV used the full 309-report cohort.

Table 2. Per-fold metrics. Mean across folds is the headline number used in §3.1; Fold 5 collapses to all-positive prediction (Specificity 0).

Fold	AUROC	AUPRC	Acc.	F1
1	0.925	0.981	0.937	0.961
2	0.936	0.978	0.927	0.954
3	0.955	0.988	0.867	0.907
4	0.987	0.997	0.908	0.939
5	0.749	0.947	0.849	0.918
mean	0.910	0.978	0.897	0.936
sd	0.083	0.017	0.034	0.020

A separate *layer-wise probing* analysis trained a linear classifier on the pooled sentence embedding from each frozen hidden layer to verify that diagnosis-relevant signal is concentrated in upper layers. The probe peaks at Layer 9 with mean AUC 0.894 (95% CI [0.785, 0.940]) and degrades toward both ends (Table 3), consistent with mid-to-upper transformer layers carrying clinical-semantic content.

Table 3. Layer-wise probing AUC (linear classifier on each frozen layer, five-fold CV). The end-to-end fine-tuned model in Table 2 unsurprisingly outperforms any single frozen-layer probe.

Layer	Mean AUC	95% CI	Mean Acc.
1	0.703	[0.587, 0.868]	0.801
2	0.741	[0.632, 0.876]	0.801
3	0.750	[0.627, 0.872]	0.801
4	0.767	[0.648, 0.883]	0.810
5	0.772	[0.644, 0.876]	0.810
6	0.844	[0.749, 0.915]	0.833
7	0.846	[0.744, 0.915]	0.845
8	0.881	[0.775, 0.931]	0.872
9	**0.894**	**[0.785, 0.940]**	**0.896**
10	0.878	[0.779, 0.929]	0.896
11	0.872	[0.772, 0.926]	0.890
12	0.858	[0.776, 0.931]	0.867

B. Silver-vs-Gold Report-Vector Validation

We validated the Llama-3.1-8B silver labels against expert-gold annotations on 15 hand-labeled reports (1,199 sentences). For each report we constructed an 18-domain proportion vector from silver labels and from gold labels, then compared them at three levels.

Five-vote self-consistency. Because the five votes come from one model rather than independent human annotators, we report them as a self-consistency stability check rather than inter-annotator reliability. Across all 25,105 high-attention sentences the five-vote Fleiss κ over the 18 domains plus the abstention category was 0.56. Votes were unanimous on 34.7% of sentences and reached a three-vote majority on a further 50.2%, while the remaining 15.1% had no majority and fell back to other-general. Mean agreement with the modal label on sentences that received a domain was 0.83. Self-consistency was therefore moderate and stable overall, with the residual disagreement concentrated in the peripheral domains flagged below.

Vector-level fidelity. Mean per-report cosine similarity between silver and gold vectors was 0.78, range 0.57 to 0.98. Pooled Spearman correlation across all 270 (domain $\times$ report) cells was 0.77.

Per-domain fidelity. ASD-core domains showed strong agreement (Table 4): insistence on sameness, restricted interests, stereotyped behavior, and nonverbal communication all reached Pearson ≥ 0.93. Peripheral psychiatric and format domains were less reliable, notably externalizing (-0.07) and internalizing (0.20). Claims about these specific domains in the modes analysis should therefore be treated with caution.

Mode and Mahalanobis robustness. Hard GMM mode assignment agreed in 10 of 15 reports (67%). Mahalanobis ranking between silver and gold was weakly anti-correlated (Spearman -0.12), and tail-membership at the 90th-percentile cutoff agreed in 11 of 15 reports.

Implication. The phenotype framework finds reasonable support for ASD-core domains at the report-vector level. Mode-level individual classifications and peripheral-domain signals carry greater silver-labeling uncertainty and should be interpreted accordingly. The 18-domain ontology is under ongoing clinical review and refinement, so per-domain attributions may shift in future analyses.

C. Multi-Seed Stability and Permutation Null

We refit GMM($k = 4$) for 50 random seeds (and a separate set of 50 bootstrap resamples) and computed pairwise Adjusted Rand Index (ARI) for all 1,225 pairs.

Cohen's d between observed (50-seed) and null is 7.82 ($p < 0.001$). Per-cluster Hungarian-matched persistence vs. seed=42: M4 1.00, M2 0.99, M1 0.98, M3 0.95. The M3 diffuse mode is the least seed-stable, consistent with its broad-cloud geometry that lacks a sharp directional anchor. We set $k = 4$ by weighing fit against this cross-seed stability, not as a uniquely optimal value.

Table 4. Per-domain silver-vs-gold proportion correlation across 15 hand-labeled reports. A dash indicates insufficient variance for correlation. Cells aggregated from silver-full vs gold paired vectors across all 15 reports.

Code	Domain	Pearson	Spearman
A1	social-emotional reciprocity	0.90	0.70
A2	nonverbal communication	0.94	0.90
A3	relationship play	0.67	0.74
A4	stereotyped behavior	0.95	0.95
A5	insistence on sameness	0.96	0.97
A6	restricted interests	0.93	0.93
A7	sensory processing	0.77	0.84
B1	externalizing	-0.07	-0.14
B2	internalizing	0.20	0.48
B3	adhd	0.68	0.67
B4	physiological function	–	–
B5	adaptive behavior	0.66	0.70
B6	intelligence/learning	0.95	0.74
B7	executive function	0.64	0.65
B8	motor skills	0.90	0.71
B9	family environment	-0.38	-0.41
C1	test scores	0.13	0.05
C2	other/general	-0.50	-0.50

Table 5. Multi-seed and bootstrap stability against permutation null.

Condition	ARI mean	95% CI	n pairs
50-seed (no bootstrap)	0.907	[0.49, 1.00]	1225
50-bootstrap resamples	0.497	[0.26, 0.76]	1225
1,000-permutation null	0.000	$[-0.011, 0.015]$	1000

D. Within-Mode Variance Decomposition

We partitioned the ASD phenotype variance into a between-mode and a within-mode part to measure how much a hard mode label leaves unexplained. On the ILR coordinates of the 248 ASD reports, a one-way multivariate partition (total sum of squares = between + within; η^2 = between/total) uses the step-5 mode assignments. In the full 17-dimensional ILR space the mode label accounts for 24.7% of the variance; in the leading three PCs where the GMM is fit it accounts for 54.2% (Table 6). A label-permutation null (5,000 shuffles, group sizes preserved) puts η^2 at 1.2% ($p < 0.001$), so mode membership separates reports above chance but leaves most variance inside the modes. Per-mode dispersion (mean squared distance to the mode centroid in the leading PCs) is largest for M3 at 116.9 and smallest for the prototype modes M1 (31.4) and M2 (26.6), with M4 at 76.8, tracking the diffuse-vs-central geometry in Table 1. The leading-PC η^2 is measured in the space where the GMM is fit and so is an upper bound; the 24.7% full-ILR value is the rotation-invariant figure. The eigenvalue spectrum of the same within-ASD ILR coordinates gives an effective dimensionality of 12.4 of 17 by an entropy measure, and 13.2 after report type is regressed out, so the representation does not collapse onto a few axes.

Table 6. Variance explained by hard mode membership (η^2) and the within-mode residual, with a label-permutation null.

Space	η^2 (between)	within	null η^2
Leading 3 PCs	54.2%	45.8%	1.2%
Full 17-dim ILR	24.7%	75.3%	1.2%

E. Exemplar High-Attention Content for the Fingerprint Case

The Section 4 exemplar is the prototype-mode (M1) patient whose fingerprint reads "prototype mode, high confidence, with within-mode tail deviation on insistence on sameness" (GMM posterior 0.95 for M1; Mahalanobis 5.5; within-M1 $z = 3.0$). Its ten highest-attention sentences are selected out-of-fold and are the text the LLM maps to the 18 domains.

Privacy. We do not reproduce verbatim report text. Table 7 gives de-identified clinical-concept keywords for the ten highest-attention sentences and the silver-label domain each one received. No names, dates, or identifying details are included.

Domain content. Six of the ten sentences received ASD-core labels: social-emotional reciprocity, nonverbal communication, relationship play, stereotyped behavior, restricted interests, and insistence on sameness. The highest-attention insistence-on-sameness sentence describes a fixed color-ordering routine and distress when the order is broken; the silver labeler assigned it insistence on sameness at confidence 0.95. The fingerprint flags this same domain as the patient's tail deviation, and it has the highest silver-vs-gold agreement in our validation (Pearson 0.96; Appendix B). The other four sentences received language (intelligence/learning, two sentences), internalizing, and one low-agreement other/general label, so the high-attention pool is not uniformly ASD-core. For one patient this supports the flagged tail deviation as clinically coherent content, but does not validate the deviation layer across patients.

F. Report-Type Confound

Reports came from two documentation streams, autism diagnostic reports and psychological assessment reports. The first principal component of the ASD phenotype space aligned with this distinction, with a point-biserial correlation of 0.84, and 13 of 18 domains differed by report type. One mode, M4, consisted almost entirely of psychological assessment reports, with a Cramér's V of 0.96 between mode and report type, and its source documents were about a quarter the length of the documents in the other modes.

We regressed report type out of each isometric log-ratio coordinate and repeated the analyses on the residuals. A Gaussian mixture still improved BIC over a single multi-

Table 7. De-identified clinical-concept keywords for the ten highest-attention sentences of the Section 4 exemplar (M1 patient), with the silver-label domain each received. $\star$ marks the insistence-on-sameness sentence the fingerprint flags as the within-mode tail deviation. Keywords are abstracted glosses; no verbatim report text is reproduced.

Rank	De-identified keyword (gloss)	Silver-label domain
1	aversion to nail/grooming care	other/general
2	caregiver-reported language delay	intelligence/learning
3	crying, anxiety when not reassured	internalizing
4	fixed color-ordering routine; distress when order broken $\star$	insistence on sameness
5	mostly single-word speech	intelligence/learning
6	poor name-response and eye contact	social-emotional reciprocity
7	follows play-scheme but over-absorbed in one toy	relationship play
8	mostly flat facial affect	nonverbal communication
9	hand-flapping; visual self-stimulation	stereotyped behavior
10	little spontaneous interest in toys	restricted interests

variate t, with a ΔBIC of -54 at $k = 4$. The residual dimensions PC4 to PC17 stayed sharply leptokurtic, with a Mardia z of 37.6 after control and 39.7 before. The four-mode partition no longer reproduced across random seeds once report type was removed, where cross-seed ARI fell to about 0.6 for $k \geq 3$ and only a two-way split stayed near 0.9. Restricting the analysis to the autism diagnostic reports alone gave the same pattern. The heavy-tailed individual deviation is therefore independent of documentation format, whereas the finer mode structure is weak and partly tied to it.