A Human-AI Collaboration System for ADHD Assessment from Primary School Reports

Florian Onur Kuhlmeier*
human-centered systems lab (h-lab)
Karlsruhe Institute of Technology
Karlsruhe, Germany
florian.kuhlmeier@kit.edu

Adrian Wegener*
Karlsruhe Institute of Technology
Karlsruhe, Germany
adrian.wegener@kit.edu

David Schulmeister
Karlsruhe Institute of Technology
Karlsruhe, Germany
david.schulmeister@student.kit.edu

Johanna Waltereit
Department of Child and Adolescent Psychiatry
LWL-Klinikum Marsberg
Marsberg, Germany
johanna.waltereit@lwl.org

Robert Waltereit
Department of Child and Adolescent Psychiatry
LWL-Klinikum Marsberg
Marsberg, Germany
robert.waltereit@lwl.org

Alexander Maedche
human-centered systems lab (h-lab)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
alexander.maedche@kit.edu

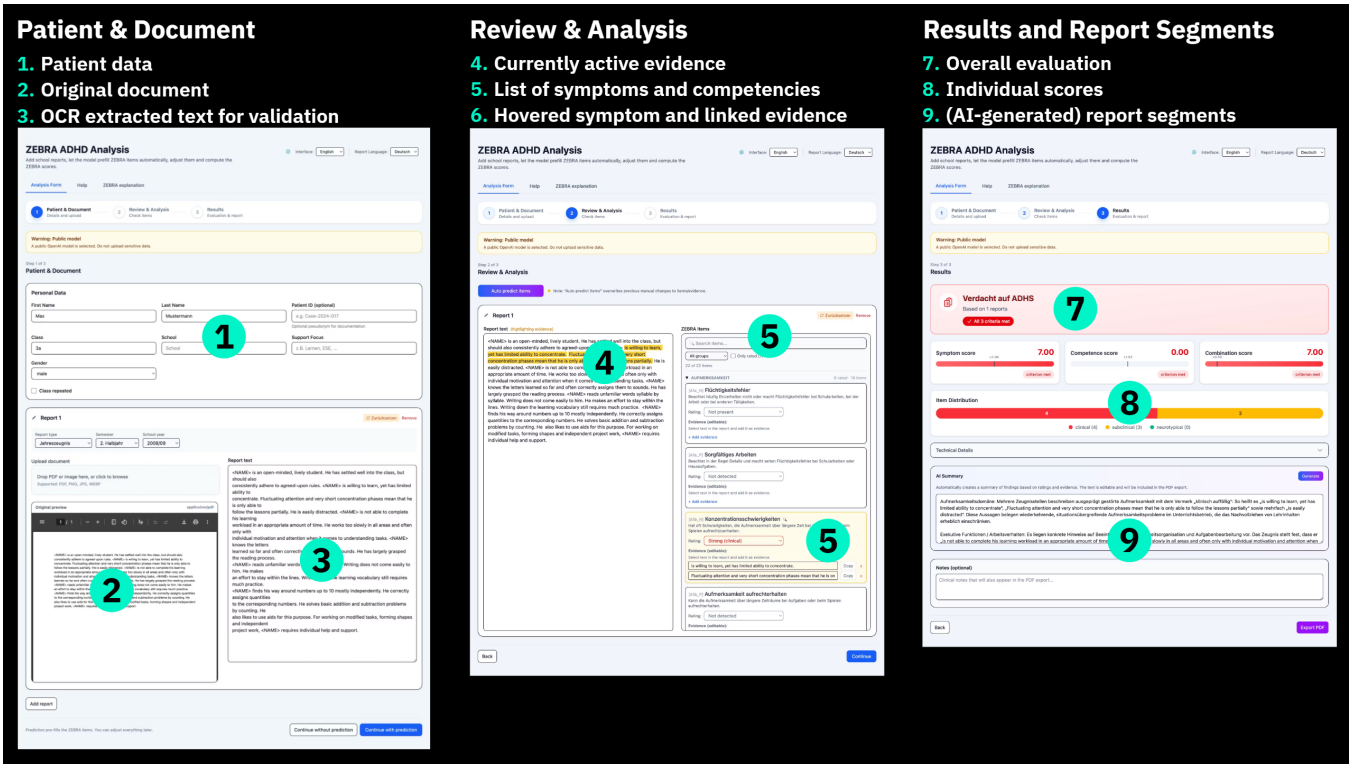

Figure 1: The workflow for analyzing primary school report text for ADHD diagnosis supported by the Human-AI Collaboration System: (1-3) Patient data entry and document upload with OCR extraction; (4-6) Side-by-side review with evidence highlighting and editable item ratings from the validated ADHD ZEBRA instrument; (7-9) Results showing the indication of ADHD diagnosis, computed symptom and competency scores, AI-generated summary, and PDF export.

Abstract

Diagnosing ADHD in adolescents and adults requires evidence of childhood symptoms, yet retrospective recall is prone to memory

*Both authors contributed equally.

CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3798770

bias. German primary school reports contain detailed narrative teacher assessments of academic and social behavior, providing contemporaneous observations. However, systematic and detailed analysis of these free-text narratives is too time-consuming for routine care. We present a Human–AI collaboration system that augments the existing validated paper-based ADHD ZEBRA diagnosis instrument for school reports, shown in Figure 1 by (1-3) extracting text from scanned reports, (4-6) proposing criterion-level symptom and competency ratings with linked evidence excerpts, and (7-9) generating summary scores and documentation-ready report segments after clinician verification. A formative evaluation with clinicians ($n = 3$) identified adoption prerequisites: evidence provenance, rule-conform instrument use, and outputs supporting clinical documentation and patient communication. Participants also described case-contingent automation needs, motivating adaptive interaction modes from rapid screening to full review. These findings suggest that human-AI collaboration systems in mental health assessment should prioritize auditability and clinician agency over automation.

CCS Concepts

• **Human-centered computing → User studies**; **Collaborative interaction**.

Keywords

ADHD, Assessment, Diagnosis, Human-AI Collaboration, User-Journey

ACM Reference Format:
Florian Onur Kuhlmeier, Adrian Wegener, David Schulmeister, Johanna Waltereit, Robert Waltereit, and Alexander Maedche. 2026. A Human-AI Collaboration System for ADHD Assessment from Primary School Reports. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3772363.3798770

1 Introduction

Attention-Deficit/Hyperactivity Disorder (ADHD) is a lifelong condition affecting approximately 2.5% of adults worldwide [4]. Many individuals diagnosed in childhood continue to meet diagnostic criteria in adulthood (up to 70%), and timely diagnosis can enable access to effective interventions and accommodations [4]. Despite increasing awareness and available support [3, 13, 14], identifying ADHD in adolescents and adults remains challenging in everyday clinical practice. A core difficulty is the required retrospective confirmation that ADHD symptoms were already present in childhood [5]. In practice, clinicians often rely on self-report and caregiver recall. However, these accounts are vulnerable to memory decay and bias [1, 9]. Primary school reports, which include free-text information, offer a complementary source of evidence: they are contemporaneous, written observations by educational professionals across academic and social contexts [16]. A particular challenge is that German school reports employ *Zeugnissprache*—a formulaic, euphemistic register in which criticism is expressed indirectly (e.g., "tries hard" to signal underperformance, "not always" to indicate inconsistency)—requiring contextual interpretation beyond simple keyword matching. Waltereit et al. proposed and validated the ADHD ZEBRA instrument, which maps free-text school report statements to ADHD symptom criteria and adaptive competencies and provides validated scoring cut-offs [16]. Despite its validity, the paper-based workflow remains labor-intensive (often 15–30 minutes per school report), limiting adoption under tight clinical schedules.

We present a web-based human–AI collaboration system that augments ADHD ZEBRA analysis while preserving clinician accountability. The system (1) extracts text from scanned school reports via OCR, (2) proposes criterion-level ratings for symptoms and competencies with linked evidence excerpts, and (3) computes ADHD ZEBRA summary scores and compares them to validated cut-offs after clinician verification. Rather than aiming for automated diagnosis, our design treats model outputs as *draft suggestions* that must remain auditable, editable, and traceable to source text. System requirements were derived from interviews with healthcare professionals who regularly use ADHD ZEBRA ($n = 2$). We report a formative evaluation with licensed clinicians ($n = 3$) using remote think-aloud sessions to surface adoption prerequisites and workflow requirements for retrospective assessment from school-report text. With our work, we contribute:

(1) A system that transforms the paper-based ADHD ZEBRA instrument [16] into an interactive human-AI collaboration workflow, featuring AI-generated initial ratings for ADHD symptoms and adaptive competencies that clinicians verify, after which the system calculates summary scores and compares them against validated cut-off thresholds.
(2) Formative findings identifying requirements for (a) evidence provenance and verification, (b) enforcement of instrument integrity constraints during interaction, and (c) integration with clinical documentation practices.
(3) Design implications for adaptive interaction modes (quick screening vs. criterion-by-criterion review) that prioritize auditability and clinician agency over automation in mental health assessment support tools.

2 Related Work

Existing research in AI-based diagnosis of ADHD has largely focused on detection and assessment via biomarkers or behavioral sensing [6, 15, 19]. In HCI, prior work has also explored interactive and wearable systems for measuring ADHD-related behavior [8]. Our work targets a different and comparatively underexplored clinical bottleneck: retrospective evidence gathering from archival documents during adolescent/adult diagnostic workups. Diagnostic guidelines require evidence that core ADHD symptoms were present in childhood [5], yet retrospective self-report and caregiver recall are vulnerable to memory bias [1, 9]. Primary school reports provide contemporaneous observations and are recommended when available [10]. Waltereit et al. validated the ADHD ZEBRA instrument for systematically mapping school-report text to ADHD symptom criteria and adaptive competencies with validated cut-offs, but the manual workflow remains time-consuming for routine practice [16]. Research on clinical decision support emphasizes that adoption depends on workflow integration and on enabling clinicians to interrogate and correct outputs while maintaining clear accountability [2]. Complementary HCI work suggests that

eliciting concrete supporting evidence can improve the quality of judgments by reducing biased, availability-driven reasoning [18]. These insights motivate our design focus on evidence-linked suggestions, rule-aware interaction that preserves instrument validity, and documentation-ready outputs for retrospective assessment.

3 Designing a Human–AI Collaborative System for Primary School Report Analysis

3.1 The Paper-Based ADHD ZEBRA Instrument

Our system builds on the ADHD ZEBRA, a validated instrument for retrospective ADHD assessment from German primary school reports [16]. ADHD ZEBRA operationalizes ICD-10/DSM-5-aligned symptom criteria by systematically mapping teacher-written school report statements to criterion-level indicators. The instrument uses a dual-scale rating approach: a *symptom scale* with three levels (0 = criterion not mentioned, 1 = subclinical expression, 2 = clinical expression) and a *competency scale* (binary) capturing adaptive competencies described in the reports. Clinicians assign ratings per criterion based on evidence excerpts from the report text and compute summary scores that are compared against validated cut-offs [16].

3.2 User Journey and Design Requirements

To understand how the ADHD ZEBRA instrument is currently used and embedded in ADHD diagnosis, we interviewed two mental health professionals who regularly analyze primary school reports using the instrument. These interviews were exploratory in scope, aimed at establishing initial design requirements for prototyping rather than a comprehensive needs analysis, but both interviewees had direct, regular experience with ADHD ZEBRA. From these interviews, we developed the user journey shown in Figure 2.

From the user journey, we derive the following requirements for an AI-augmented ADHD ZEBRA workflow: (1) The system should support scanned and heterogeneous school reports via OCR and enable clinicians to inspect and correct extracted text. (2) The system should reduce manual mapping effort by proposing criterion-level draft ratings together with explicit links to candidate supporting excerpts in the original report. (3) The system should generate documentation-ready text segments that summarize findings in a structured way, explicitly reflecting both symptom-related evidence and adaptive competencies. (4) The system should provide fine-grained controls to verify, edit, and override AI proposals (ratings and evidence), and to complete the assessment in a rule-conform way before scores and report segments are finalized.

3.3 The Human-AI Collaboration System enabling AI-Augmented ADHD ZEBRA

These requirements motivate a Human–AI collaboration approach in which the system supports drafting and the clinician remains the final decision-maker: the AI proposes ratings and evidence excerpts, clinicians verify and adjust them, and results (scores and report segments) are produced only after verification.

Workflow. We implemented this approach as a web application with a three-step workflow (Fig. 1).

- *Step 1: Patient & Document.* Clinicians enter patient metadata and upload primary school reports as PDF/image files. The system performs OCR to extract text and displays it for review and correction.
- *Step 2: Review & Analysis (drafting + verification).* The interface presents a side-by-side view with the school report text on the left and the ADHD ZEBRA criteria on the right. For each criterion, the system displays (a) the criterion description, (b) a draft rating suggestion (symptom scale: 0/1/2; competency scale: 0/1), and (c) one or more candidate evidence excerpts linked to highlighted spans in the report text. Clinicians can validate and modify both the proposed ratings and the linked evidence excerpts.

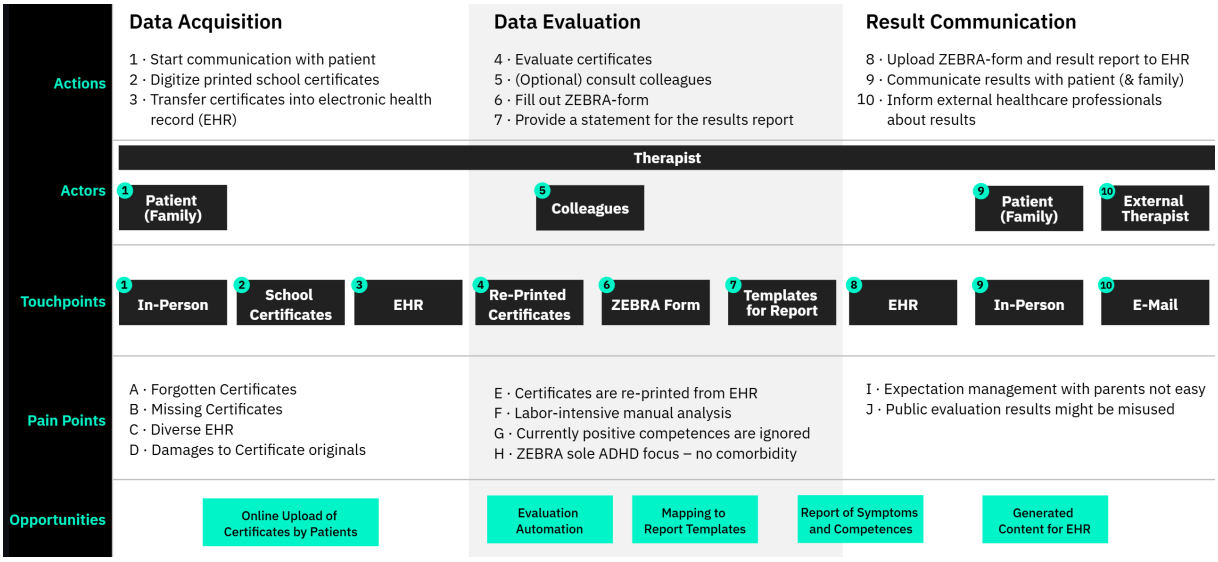

Figure 2: User journey based on interviews with healthcare professionals (n=2).

- *Step 3: Results and report segments.* After verification, the system computes the symptom, competency, and combined scores and compares them against validated cut-offs [16]. It then generates editable *report segments* organized by diagnostic domain that summarize findings based on symptoms and competencies, retaining traceability to criterion-level evidence excerpts. Finally, clinicians can export a PDF report for clinical documentation and communication.

Model Evaluation. Because primary school reports constitute sensitive personal data, data protection requirements restricted us to open-weight models hosted on-premise.

We used the ADHD ZEBRA dataset from Waltereit et al. [16] (300 students; 890 report segments; 19,590 criterion-level labels). Ground-truth ratings were produced by two experienced clinician raters blind to diagnostic status, with high inter-rater reliability (average weighted Cohen's $\kappa = 0.90$).

We compared four model approaches: a fine-tuned German encoder (TiME-de-xs, 22M parameters) [12] and three LLMs (Qwen 2.5 235B [17], GPT-OSS 120B [11], Mixtral 8x22B [7]). The dataset exhibits severe class imbalance (Table 1). We therefore treat Macro F1 as the primary metric.

Table 1: Class distribution of the symptom scale in the evaluation dataset.

Class	Count	%
0 (No symptom)	15,914	81.2%
1 (Mild)	1,636	8.4%
2 (Severe)	2,040	10.4%

The encoder was fine-tuned with focal loss and inverse-frequency class weighting to address this imbalance (details in Appendix). Table 2 summarizes performance. The encoder achieves the highest accuracy (82.2%) and recall for severe symptoms (55.5%), making it suitable when detection sensitivity is prioritized. Qwen 2.5 leads in Macro F1 (0.558) and Cohen's κ (0.536), indicating the best overall calibration across classes. Notably, the encoder shows near-zero F1 for mild symptoms, consistently misclassifying them as absent or severe, which limits its suitability for screening or monitoring cases where subtle indications matter.

Table 2: Model performance comparison on ADHD ZEBRA criterion classification. Best values per metric shown in bold. Rec/Prec refer to severe-symptom class.

Model	Acc.	Macro F1	Rec._{sev}	$\text{Prec.}_{\text{sev}}$	κ
Qwen 2.5 (235B)	71.1%	**0.558**	50.1%	61.9%	**0.536**
GPT-OSS 120B	79.7%	0.521	25.0%	**76.7%**	0.504
Mixtral 8x22B	69.4%	0.508	36.2%	60.5%	0.470
Encoder (local)	**82.2%**	0.472	**55.5%**	47.0%	0.454

4 Formative Evaluation

4.1 Methodology

We conducted a formative qualitative evaluation of the AI-augmented ADHD ZEBRA prototype to understand implications for integrating it into the routine ADHD diagnostic workflow. We used semi-structured, remote think-aloud sessions focused on interpretability, workflow fit, and clinical usefulness. The study received IRB approval. We interviewed three licensed clinicians involved in outpatient ADHD diagnostic workups ($n = 3$), intentionally sampling across roles and prior familiarity with ADHD ZEBRA. P1 (licensed psychotherapist) and P3 (psychiatrist) had no prior ADHD ZEBRA experience; P2 (licensed psychotherapist and academic researcher) reported frequent ADHD ZEBRA use. Session durations were 36:36, 1:09:00, and 32:49. Each session followed the end-to-end workflow supported by the prototype using mock patient information and an example primary school report: participants uploaded a report (including OCR review), triggered AI pre-filling, inspected and edited criterion ratings and linked evidence excerpts, reviewed computed summary scores against validated cut-offs, and exported a PDF report. Sessions concluded with reflective questions about trust, accountability, and expected integration into routine practice. The interviews were conducted by two researchers. One researcher then coded the transcripts, iteratively refined codes, merged them into themes and discussed uncertainties with the second researcher until consensus was reached.

4.2 Findings and Design Implications

We derived themes through inductive thematic analysis and report the main themes from the interviews. Participants noted substantial variation in how primary school reports are handled in routine ADHD diagnosis, from rapid skimming to systematic analysis. This variation shaped expectations of AI-based assistance under time constraints:

> *We're very tightly scheduled … I don't know where I'd find 20 minutes per patient on top.* (P1)

Importantly, the preferred degree of automation was case-contingent:

> *It depends on the individual case as some are straightforward and others aren't.* (P2)

Therefore, in straightforward cases, higher levels of automation were seen as acceptable, whereas in complex or ambiguous cases the Human–AI collaboration aspects became more important. Three themes informed design implications.

Evidence provenance, transparency, and verification. Participants valued traceability from suggested ratings to source text; for example, P3 pointed to hover-based evidence highlighting as helpful. However, provenance was only useful when paired with efficient verification and correction. In this sense, trust meant the ability to verify, correct, and document suggestions within the interface. P1 and P3, who had no prior ADHD ZEBRA experience, indicated a need for plain-language framing of what the outputs represent (e.g., where criterion definitions and cut-offs come from, how scores are computed, and what the system cannot conclude).

Instrument integrity: supporting rule-conform use under AI assistance. Participants highlighted risks of instrument-rule

violations when draft outputs are accepted uncritically, such as assigning the same excerpt to multiple criteria:

The same sentence is used for two different criteria. (P2)

This motivates interaction support that surfaces incomplete or inconsistent states during editing (e.g., missing evidence for a non-zero rating) and helps clinicians resolve them before results are finalized. Participants further noted that some criteria can be grounded more directly in observable classroom behavior than others, suggesting that the interface should help prioritize attention toward criteria that are more ambiguous and likely to require verification.

Workflow integration: adaptive interaction depth and documentation-ready outputs. Because acceptable automation depends on case complexity, a fixed interaction depth is unlikely to fit routine practice. Participants' accounts motivated two complementary interaction modes: a *Quick-Assessment* mode for rapid review of summary scores and key evidence excerpts when cases are straightforward, and a *Full-Review* mode for criterion-by-criterion verification and evidence editing when cases are complex or contested. This motivates a *Prioritized-Review* sub-mode within *Full-Review* that shows only criteria where model certainty falls below a threshold, limiting clinician review to items the model is insufficiently confident about while accepting higher-confidence suggestions automatically. Participants also discussed retrospective assessment as a team process in which delegated staff might prepare drafts and clinicians review and sign off, increasing the importance of traceability and clear responsibility markers. Finally, participants emphasized the PDF export as a key operational deliverable for routine care (P3), both for documentation and communication with patients and caregivers. Generated text should therefore follow clinical writing conventions and remain defensible (e.g., indirect/subjunctive phrasing rather than verbatim quotation, as noted by P1).

5 Discussion

Our formative findings suggest that effective AI augmentation for retrospective ADHD assessment from archival school reports is less about maximizing automation and more about supporting case-contingent workflows. When cases are straightforward, clinicians may accept higher automation (e.g., rapid pre-filling with minimal edits). When cases are complex, ambiguous, or contested, the value shifts toward Human–AI collaboration features that preserve clinician agency: evidence provenance (ratings linked to excerpts), efficient verification and correction, and interaction support that helps complete a rule-conform ADHD ZEBRA assessment. This aligns with prior HCI work showing that clinical decision support is adopted when it fits workflows and supports accountability rather than replacing judgment [2]. It also resonates with evidence-based prompting findings: interfaces that foreground concrete evidence can shift reasoning toward more reflective, less bias-prone judgments [18]. A second implication is that instrument integrity and documentation should be treated as important design goals. Because the ADHD ZEBRA is a validated instrument with scoring rules and cut-offs [16], AI drafting must not only propose ratings but also support rule-conform evidence assignment and defensible exports. In our setting, school reports contain sensitive personal data, which constrains practical deployments toward on-premise processing. This further motivates interaction designs that compensate for remaining model limitations through verification-centric UI patterns, explicit unresolved states, and documentation-ready report segments grounded in symptoms and competencies. The encoder's near-zero performance on mild symptoms (Class 1) is particularly consequential for this design challenge: missed mild indicators may go unnoticed if clinicians over-rely on AI pre-filling, while a high rate of false negatives could erode trust over time. The *Full-Review* mode, which requires criterion-by-criterion verification with explicit evidence links, is designed to mitigate this risk; however, the resulting cognitive overhead—especially in complex cases where mild signals are most relevant—warrants targeted investigation in future evaluations.

5.1 Limitations

First, the proposed system is a prototype and the presented evaluation is formative. We conducted remote think-aloud sessions with $n = 3$ clinicians and focused on workflow requirements rather than diagnostic accuracy or time savings. The prototype was evaluated with example materials rather than a longitudinal deployment integrated into clinic IT systems. Second, the approach is currently tailored to German primary school reports and their characteristic *Zeugnissprache*. The ADHD ZEBRA instrument itself is specific to Germany. Adaptation to other educational systems or languages would require retraining models on new data and localizing prompt design; however, the underlying human-AI workflow—OCR, AI-drafted criterion ratings, and clinician verification—is language-agnostic in principle and may generalize given appropriate data. Third, in routine care, differential diagnosis is often the key question. However, our current system is limited to ADHD. Extending the approach to support structured analysis of signals relevant to common differential diagnoses remains future work. Finally, model performance remains limited, reinforcing that outputs should be treated as suggestions requiring clinician verification.

5.2 Next Steps

Next, we will implement adaptive interaction modes that reflect case-contingent automation: a *Quick-Assessment* mode for rapid triage using summary scores and key evidence, a *Full-Review* mode for criterion-by-criterion verification and rule-conform editing, and a *Prioritized-Review* mode that surfaces only criteria where model certainty falls below a configurable threshold for clinician review. To improve model performance, we will explore combining predictions from multiple models and multi-agent workflows in which one model proposes ratings and another critiques them before presenting results to the clinician. We will also expand the dataset with additional school reports to improve performance on rare criteria and mild symptoms, where current training data is sparse. We will then conduct a larger evaluation study with clinicians involved in diagnosing ADHD to assess workflow fit and perceived usefulness at scale, and to quantify efficiency and inter-rater agreement relative to the paper-based ADHD ZEBRA procedure and our prototype without AI-augmentation. We will also evaluate role-based delegation workflows in which clinical assistants prepare the primary school report analysis that the clinician in charge of the diagnostic procedure then reviews.

References

[1] V. Breda, L. A. Rohde, A. M. B. Menezes, L. Anselmi, A. Caye, D. L. Rovaris, E. S. Vitola, C. H. D. Bau, and E. H. Grevet. 2020. Revisiting ADHD age-of-onset in adults: to what extent should we rely on the recall of childhood symptoms? *Psychological Medicine* 50, 5 (April 2020), 857–866. doi:10.1017/S003329171900076X

[2] Eleanor R. Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne Currie, J. Marc Overhage, Erika S Poole, and Jofish Kaye. 2023. Healthcare AI Treatment Decision Support: Design Principles to Enhance Clinician Adoption and Trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–19. doi:10.1145/3544548.3581251

[3] Jaehyun Byun, Chowon Joung, Yerim Lee, Suyun Lee, and Wooky Won. 2025. Le Petit Care: A Child-Attuned Design for Personalized ADHD Symptom Management Through AI-powered Extended Reality. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3706599.3720300

[4] Samuele Cortese, Mark A. Bellgrove, Isabell Brikell, Barbara Franke, David W. Goodman, Catharina A. Hartman, Henrik Larsson, Frances R. Levin, Edoardo G. Ostinelli, Valeria Parlatini, Josep A. Ramos-Quiroga, Margaret H. Sibley, Anneka Tomlinson, Timothy E. Wilens, Ian C.K. Wong, Nina Hovén, Jeremy Didier, Christoph U. Correll, Luis A. Rohde, and Stephen V. Faraone. 2025. Attention-deficit/hyperactivity disorder (ADHD) in adults: evidence base, uncertainties and controversies. *World Psychiatry* 24, 3 (2025), 347–371. doi:10.1002/wps.21374 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wps.21374.

[5] Stephen V. Faraone, Mark A. Bellgrove, Isabell Brikell, Samuele Cortese, Catharina A. Hartman, Chris Hollis, Jeffrey H. Newcorn, Alexandra Philipsen, Guilherme V. Polanczyk, Katya Rubia, Margaret H. Sibley, and Jan K. Buitelaar. 2024. Attention-deficit/hyperactivity disorder. *Nature Reviews Disease Primers* 10, 1 (Feb. 2024), 11. doi:10.1038/s41572-024-00495-0

[6] Runqing Gao, Kesui Deng, and Miaoyun Xie. 2022. Deep learning-assisted ADHD diagnosis. In *Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences*. ACM, Amsterdam Netherlands, 142–147. doi:10.1145/3570773.3570849

[7] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. arXiv:2401.04088 [cs.LG] https://arxiv.org/abs/2401.04088

[8] Xinlong Jiang, Yiqiang Chen, Wuliang Huang, Teng Zhang, Chenlong Gao, Yunbing Xing, and Yi Zheng. 2020. WeDA: Designing and Evaluating A Scale-driven Wearable Diagnostic Assessment System for Children with ADHD. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. doi:10.1145/3313831.3376374

[9] Carlin J. Miller, Jeffrey H. Newcorn, and Jeffrey M. Halperin. 2010. Fading Memories: Retrospective Recall Inaccuracies in ADHD. *Journal of Attention Disorders* 14, 1 (July 2010), 7–14. doi:10.1177/1087054709347189

[10] National Institute for Health and Care Excellence (Great Britain). 2019. *Attention deficit hyperactivity disorder: diagnosis and management*. Number NG87 in NICE guideline. National Institute for Health and Care Excellence, London.

[11] OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. arXiv:2508.10925 [cs.CL] https://arxiv.org/abs/2508.10925

[12] David Schulmeister, Valentin Hartmann, Lars Klein, and Robert West. 2025. TiME: Tiny Monolingual Encoders for Efficient NLP Pipelines. arXiv:2512.14645 [cs.CL] https://arxiv.org/abs/2512.14645

[13] Lucas M. Silva, Franceli L. Cibrian, Clarisse Bonang, Arpita Bhattacharya, Aehong Min, Elissa M Monteiro, Jesus Armando Beltran, Sabrina Schuck, Kimberley D Lakes, Gillian R. Hayes, and Daniel A. Epstein. 2024. Co-Designing Situated Displays for Family Co-Regulation with ADHD Children. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3613904.3642745

[14] Evropi Stefanidi, Johannes Schöning, Yvonne Rogers, and Jasmin Niess. 2023. Children with ADHD and their Care Ecosystem: Designing Beyond Symptoms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3544548.3581216

[15] Qurat Ul Ain, Soyiba Jawed, Ahmad Rauf Subhani, Wasi Haider Butt, and Muhammad Usman Akram. 2025. Examining AI-Powered ADHD Diagnosis: Current Trends, Key Challenges, and Future Directions in the Field. *IEEE Access* 13 (2025), 93148–93177. doi:10.1109/ACCESS.2025.3567427

[16] Johanna Waltereit, Martin Schulte-Rüther, Veit Roessner, and Robert Waltereit. 2024. Retrospective assessment of ICD-10/DSM-5 criteria of childhood ADHD from descriptions of academic and social behaviors in German primary school reports. *European Child & Adolescent Psychiatry* (July 2024). doi:10.1007/s00787-024-02509-4

[17] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024).

[18] Junti Zhang, Zicheng Zhu, Jingshu Li, and Yi-Chieh Lee. 2025. Mining Evidence about Your Symptoms: Mitigating Availability Bias in Online Self-Diagnosis. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–23. doi:10.1145/3706598.3713805

[19] Cuijie Zhao, Yan Xu, Ruixing Li, Huawei Li, and Meng Zhang. 2025. Artificial intelligence in ADHD assessment: a comprehensive review of research progress from early screening to precise differential diagnosis. *Frontiers in Artificial Intelligence* 8 (Sept. 2025), 1624485. doi:10.3389/frai.2025.1624485