PRISM: PATIENT-SPECIFIC REASONING AND INFORMATION SEEKING FOR MULTIMODAL PSYCHIATRIC DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal reasoning under partial observability requires acquiring evidence whose decision value justifies its cost. However, predefined prediction pipelines provide limited support for adapting evidence sources and analysis operations to evolving information needs. We formulate psychiatric diagnosis from electronic health records (EHR) and electroencephalography (EEG) as a patient-specific reasoning and information-seeking problem and propose **PRISM**, a large language model (LLM) agent framework for joint source and analysis selection. First, we construct **PsyNexus**, a longitudinal multimodal cohort of **2,051 patients** with major depressive disorder, bipolar disorder, or schizophrenia, comprising 2,799 encounter-linked EEG recordings and 58,514 EHR encounters. The encounter linkage supports interpreting neural observations within each patient's clinical context. Second, we develop PRISM-Base, an evidence-grounded agent that maintains a patient-level diagnostic state and integrates clinical facts, explicit EEG measurements, and optional learned representations through source-linked tools. Third, we introduce **P**rogressive **E**vidence **S**eeking and **R**easoning (**PESR**), a reinforcement learning alignment strategy for resource-efficient evidence acquisition and analysis. PESR evaluates diagnostic refinement with a frozen assessor and progressively introduces evidence grounding, information value, and cost sensitivity, allowing informative acquisition to develop before optimizing resource expenditure. On PsyNexus, PRISM achieves an absolute macro-F1 improvement of 5.1% while reducing inference time by 60% relative to full-evidence multimodal supervised fine-tuning. This work provides a methodological foundation for adaptive multimodal intelligence, supporting the development of evidence-seeking agents that reason and make decisions under uncertainty and resource constraints.

1 INTRODUCTION

In sequential decision-making, the value of additional evidence depends on the observations already acquired (Foster et al., 2021). Multimodal agents must therefore determine which sources and analyses can inform the current decision. Psychiatric diagnosis provides a setting for studying this dependence. Electronic health records (EHR) capture longitudinal clinical histories (Shmatko et al., 2025), while electroencephalography (EEG) provides neurophysiological observations (Wu et al., 2020; Li et al., 2025). Neuroimaging studies of major depressive disorder (MDD), bipolar disorder (BD), and schizophrenia (SCZ) identify cross-disorder overlap and substantial individual heterogeneity (Hettwer et al., 2022; Segal et al., 2023). These findings motivate interpreting neural observations within each patient's clinical context. Accordingly, measurement value should be assessed conditional on available evidence, motivating acquisition and analysis that adapt to the evolving patient state (Figure 1).

Recent work has advanced both evidence representation and active acquisition. EEG foundation models such as CBraMod (Wang et al., 2025) and REVE (El Ouahidi et al., 2025) provide transferable signal representations. EDDI (Ma et al., 2019) and ACO (Valancius et al., 2024) formulate feature acquisition in terms of information value and cost. LA-CDM (Bani-Harouni et al., 2026) uses diagnostic hypotheses and confidence to select clinical tests, while CDEG (Dai et al., 2026)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

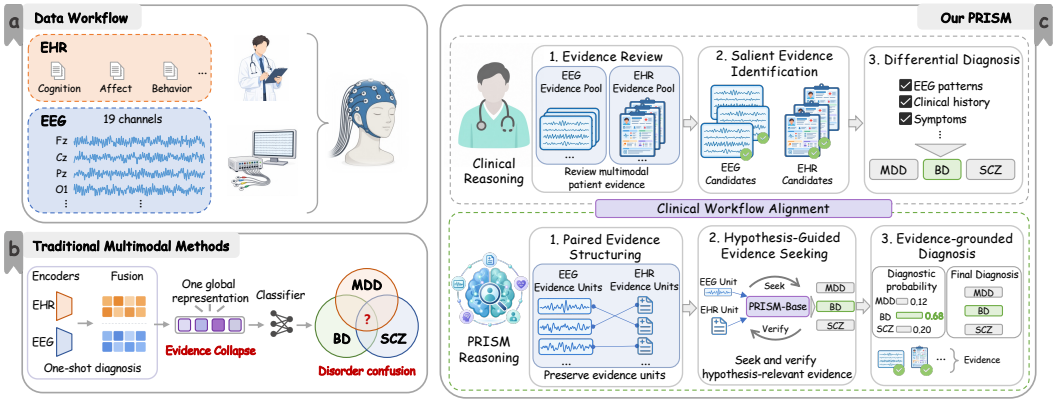


Figure 1: Overview of PRISM. (a) EHR and 19-channel EEG. (b) Conventional multimodal fusion for single-stage prediction. (c) PRISM integrates evidence structuring, targeted information seeking, and evidence-grounded diagnosis.

guides evidence acquisition and reappraisal. These approaches establish active information seeking as a component of diagnostic reasoning. With raw EEG, however, retrieving a source leaves its analysis to be determined. Frequency bands, connectivity estimators, channel groups, and learned representations provide different evidence views with distinct computational requirements. Consequently, **source selection** must identify informative records or signal intervals, while **analysis selection** must determine which measurements or representations to derive. Their joint value depends on acquired clinical and neural evidence, requiring a policy that coordinates both decisions under resource constraints.

To address this challenge, we propose **PRISM**, a large language model (LLM) agent framework for joint source and analysis selection. First, we construct **PsyNexus**, a longitudinal cohort of 2,051 patients with MDD, BD, or SCZ, comprising 2,799 encounter-linked EEG recordings and 58,514 EHR encounters. This linkage situates neural evidence within each patient's clinical course. Next, we develop PRISM-Base through supervised fine-tuning (SFT) to integrate clinical facts, explicit EEG measurements, and optional learned representations through source-linked tools. Its evolving patient state guides subsequent acquisition, analysis, and reporting. Finally, we introduce **P**rogressive **E**vidence **S**eeking and **R**easoning (**PESR**), a reinforcement learning (RL) alignment strategy informed by sequential experimental design (Foster et al., 2021). A frozen assessor evaluates diagnostic refinement, while progressive training introduces evidence grounding and information value before cost sensitivity. This ordering enables the policy to learn informative acquisition and analysis before optimizing resource expenditure. Our contributions are threefold:

- We propose **PRISM**, an LLM agent framework that jointly selects evidence sources and neural analyses. Its **PESR** strategy progressively aligns evidence-seeking policies with conditional information value and resource cost, advancing active multimodal reasoning under partial observability.
- **PsyNexus** provides a longitudinal cohort linking clinical histories with encounter-level EEG across three psychiatric disorders. This alignment situates neural observations within patient-specific clinical contexts, supporting research on interpretable multimodal reasoning and adaptive evidence acquisition.
- Experiments demonstrate a 5.1% absolute macro-F1 gain and 60% lower latency than full-evidence multimodal SFT. Controlled analyses support joint source and analysis selection and progressive alignment for resource-efficient multimodal decision-making.

2 Related Work

Multimodal medical models and diagnostic agents. MMed-RAG (Xia et al., 2025), MedAgent-Pro (Wang et al., 2026b), and multimodal AMIE (Saab et al., 2026) extend medical reasoning

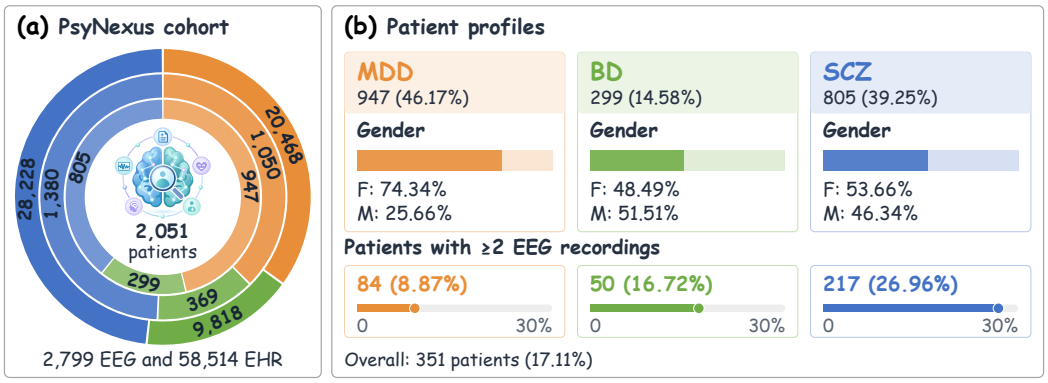


Figure 2: PsyNexus cohort and patient profiles. (a) Inner, middle, and outer rings show patients, EEG recordings, and EHR encounters. (b) Gender composition (F for female and M for male) and patients with at least two EEG recordings. All EEG recordings are encounter-linked.

through retrieval and tools. LA-CDM (Bani-Harouni et al., 2026) learns clinical test selection; ProMed (Ding et al., 2026) and CDEG (Dai et al., 2026) evaluate evidence utility. ReTool (Feng et al., 2026) uses outcome feedback, while EmotionThinker (Wang et al., 2026a) progressively introduces reasoning rewards. PRISM extends acquisition to signal-analysis decisions, using a frozen utility assessor and progressive rewards to balance conditional evidence value with cost.

Multimodal psychiatric diagnosis. EEG connectivity reveals shared and individualized psychosis patterns (Li et al., 2025), while EEG and functional near-infrared spectroscopy support bipolar classification (Tahir et al., 2026). AgentMental (Hu et al., 2026) studies adaptive depression interviews, and CBraMod (Wang et al., 2025) and REVE (El Ouahidi et al., 2025) provide EEG representations. PRISM combines explicit measurements and learned representations with longitudinal EHR to learn context-dependent diagnostic value without assuming disease-specific connectivity signatures (Appendix A).

3 METHOD

PRISM combines diagnostic reasoning, source selection, and neural analysis (Figure 3). PsyNexus organizes longitudinal EHR and encounter-linked EEG into accessible sources. PRISM-Base, initialized from Qwen3.5-9B (Qwen Team, 2026), learns evidence interpretation and tool interaction. We adopt an open-weight backbone to process sensitive patient data within a controlled local environment, avoiding transmission to externally hosted model services. PESR then aligns trajectories with conditional information value and resource cost. Appendix B provides implementation details.

3.1 PSYNEXUS: COHORT AND EVIDENCE ORGANIZATION

Overall information. PsyNexus contains 2,051 patients (947 MDD, 299 BD, 805 SCZ), 58,514 EHR encounters, and 2,799 encounter-linked EEG recordings. It is the largest cohort in Table 1, combining longitudinal clinical assessments, treatments, tests, vital signs, and psychometric measures with EEG for three-disorder diagnosis. Repeated linked EEG recordings are available for 351 patients (Figure 2; Appendices B.1 and B.2). We will make the processed data publicly available for non-commercial academic research.

Diagnostic episodes. Each episode uses evidence available at an index encounter to retrospectively identify $y \in \{\text{MDD}, \text{BD}, \text{SCZ}\}$. Longitudinal diagnostic review determines the reference independently of the input cutoff. Episodes specify initial evidence and an admissible clinical and EEG archive. Target-revealing content is masked, and subsequent label-review records remain inaccessible. All encounters and trajectories of a patient stay in one partition. Appendix B.3 details label review and access boundaries; Section 4.1 describes evaluation.

Table 1: Comparison of PsyNexus with psychiatric EEG resources. N denotes participants. ✓ marks data available for at least a subset, and ✗ denotes absence from the examined release.

Dataset	N	Labels	Evidence availability						
			Raw EEG	Notes	Scales	Tx	Tests	Long. EHR	Link
TDBRAIN (van Dijk et al., 2022)	1,274	Multi.	✓	✗	✓	✓	✗	✗	✗
MODMA (Cai et al., 2022)	53	MDD, HC	✓	✗	✓	✗	✗	✗	✗
BRMH (Park et al., 2021)	945	Multi.	✗	✗	✓	✗	✗	✗	✗
FEP Resting EEG (Salisbury et al., 2022)	61	FEP, HC	✓	✗	✓	✓	✗	✗	✗
RepOD (Olejarczyk & Jernajczyk, 2017)	28	SCZ, HC	✓	✗	✗	✗	✗	✗	✗
PsyNexus (ours)	**2,051**	MDD, BD, SCZ	✓	✓	✓	✓	✓	✓	✓

Multi.: multiple diagnoses or referral indications. *HC*: healthy controls. *FEP*: first-episode psychosis. *Notes*: narrative clinical records. *Scales*: psychometric assessments. *Tx*: treatment information. *Tests*: laboratory results and clinical examination reports. *Long. EHR*: longitudinal EHR. *Link*: EEG-to-EHR encounter mapping. MODMA denotes the 128-channel subset. Repeated EEG or scales alone do not establish longitudinal EHR (Appendix B.1).

3.2 Problem Formulation

We consider a partially observed patient archive at a fixed diagnostic time. The patient state and diagnostic distribution are

$$s_t = (O_t, E_t, p_t, \mathbf{b}_t), \qquad p_t = D_\theta(E_t) \in \Delta^2, \tag{1}$$

where t indexes decisions, O_t records accesses and operations, E_t contains acquired evidence, and $\mathbf{b}_t$ is the remaining budget. D_θ extracts policy class probabilities at a fixed diagnostic prefix; Δ^2 is the three-disorder probability simplex. The policy π_θ selects $a_t \in \mathcal{A}(s_t)$ subject to source availability, prerequisites, and budget. Acquisition returns evidence e_t at cost $\mathbf{c}_t$; termination produces the final report. These operations form trajectory τ. We optimize

$$\max_\theta \ \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)], \qquad \sum_{t=0}^{T-1} \mathbf{c}_t \preceq \mathbf{b}_0, \tag{2}$$

where θ denotes policy parameters, R is defined in Section 3.5, T counts acquisition and analysis operations, and $\preceq$ is componentwise comparison. Following EDDI (Ma et al., 2019) and ACO (Valancius et al., 2024), these sequential decisions balance diagnostic value against archive-access and computation costs, with a separate reserve for reporting.

3.3 Source-Linked Clinical and Neural Evidence

Clinical evidence. Retrieval returns clinical text or structured fields with source identifiers, assertion status, event time, and availability. These attributes preserve clinical context and enable traceable citations. Missing observations remain unknown, and medication orders retain their status as prescriptions.

Explicit neural measurements. For an accessible EEG interval, we construct

$$\mathcal{M}_j \in \mathbb{R}^{B\times R\times C\times C}, \qquad \mathcal{M}_j[b, r, u, v] = \Phi_r(x_{j,u}, x_{j,v} \mid b), \tag{3}$$

where j indexes the recording and interval, $x_{j,u}$ is channel u, Φ_r is estimator r, and u, v index channels. We use $B = 5$ bands $(\delta, \theta, \alpha, \sigma, \beta)$, $C = 19$ channels, and $R = 4$ measures: coherence, pairwise phase consistency (PPC) (Vinck et al., 2010), debiased squared weighted phase lag index (wPLI) (Vinck et al., 2011), and orthogonalized amplitude-envelope correlation (oAEC) (Hipp et al., 2012). Relative band power $P_j \in \mathbb{R}^{B\times C}$ complements connectivity. A measurement operation exposes requested entries or predefined channel-group summaries as structured text, retaining interval, band, estimator, channel, validity, and quality fields. This representation makes each numerical observation traceable to its analysis configuration. Invalid estimates are masked, and preprocessing and estimator settings are fixed before evaluation. All measurements describe sensor-level activity.

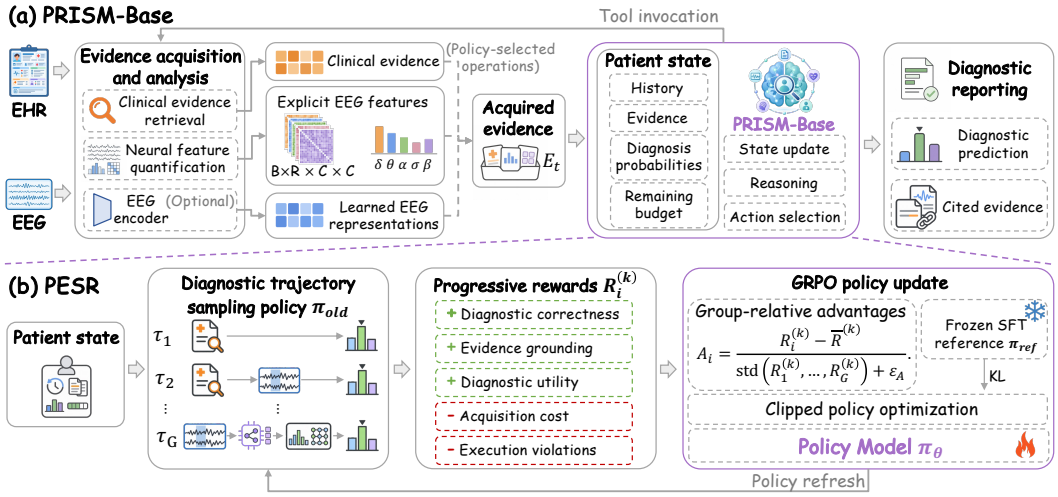


Figure 3: PRISM architecture and PESR training. (a) Acquired evidence updates the patient state before further acquisition or diagnostic reporting. History records accesses and operations. (b) Trajectories share initial evidence and budgets. A frozen assessor scores diagnostic utility, balanced against resource cost through progressive rewards. GRPO updates policy LoRA parameters with KL regularization toward the frozen SFT reference, followed by sampling-policy synchronization. Flames and snowflakes denote trainable and frozen components.

Optional learned representations. Encoding a retrieved waveform x_j yields $H_j = \mathrm{Enc}_\phi(x_j)$ with frozen ϕ. A projector maps pooled EEG tokens into the language-model embedding space alongside clinical and measurement text. We use REVE (El Ouahidi et al., 2025) and compare CBraMod (Wang et al., 2025). Tokens enter acquired evidence only when encoding is selected. They complement verifiable explicit measurements with learned signal information; encoder use is logged separately from measurement citations.

3.4 PRISM-Base

Evidence-conditioned interaction. Figure 3a integrates source retrieval, neural quantification, optional encoding, and reporting. Quantification without encoding permits selective analysis depth. Source-indexed responses update E_t and O_t; $\mathbf{b}_{t+1} = \mathbf{b}_t - \mathbf{c}_t$ and $p_{t+1} = D_\theta(E_{t+1})$ condition the next decision. Termination returns the most probable diagnosis, class probabilities, and cited evidence.

Evidence-grounded initialization. SFT uses approximately 5,000 demonstrations from 500 training patients across all three disorders. Demonstrations vary evidence combinations, trajectory lengths, and tools, with partial-evidence prefixes supporting state-conditioned supervision. Clinical fields, numerical EEG observations, tool calls, and diagnoses are verified for executability, numerical consistency, and source support. SFT updates language-model low-rank adapters (LoRA) and the EEG projector while freezing the encoder (Appendix B.5).

A fixed diagnostic prefix restricts output to three single-token disease codes. Softmax over their logits defines D_θ; the largest probability determines diagnosis. An auxiliary SFT-trained head supplies probabilities exclusively for reward-time diagnostic utility assessment.

3.5 PESR: Progressive Evidence Seeking and Reasoning

PESR uses group relative policy optimization (GRPO) (Shao et al., 2024) to align diagnostic trajectories with information value and resource cost (Figure 3b). The SFT checkpoint initializes the policy, fixed reference π_{ref}, and separate assessor $D_{\bar{\psi}}$. RL updates only policy LoRA parameters; the reference, assessor, EEG encoder, and projector remain frozen, preserving a fixed utility criterion.

Diagnostic reliability. For trajectory τ_i, prediction $\widehat{y}_i$, and label y, $R_{\text{diag},i} = \mathbb{1}[\widehat{y}_i = y]$ rewards correctness. $R_{\text{ground},i} \in [0, 1]$ scores factual support against acquired sources, including temporal and numerical attributes. Unsupported claims and empty explanations receive no grounding credit (Section 4.1).

State-dependent information value. We quantify diagnostic refinement from acquired evidence using

$$F(E, y) = -\|D_{\bar{\psi}}(E) - \mathbf{e}_y\|_2^2, \qquad U_i = \frac{F(E_{i,T_i}, y) - F(E_0, y)}{2}, \tag{4}$$

where $\mathbf{e}_y$ is the one-hot label, E_0 is initial evidence, and E_{i,T_i} is evidence after T_i operations. F is negative multiclass Brier error; $U_i \in [-1, 1]$ is its normalized improvement, a label-referenced surrogate for decision-relevant uncertainty reduction (Gneiting & Raftery, 2007). Joint clinical and neural assessment makes value conditional on prior observations. Negative gains are retained; labels enter rewards only. Appendices B.6 and B.8 detail this interpretation and matched-state analysis.

Progressive alignment. Following progressive reward introduction in EmotionThinker (Wang et al., 2026a), PESR uses

$$R_i^{(k)} = R_{\text{diag},i} + \lambda_g(k) R_{\text{ground},i} + \lambda_u(k) U_i - \lambda_c(k) C_i - \lambda_v V_i, \tag{5}$$

where k indexes updates, C_i measures normalized acquisition and processing cost, V_i counts execution violations, and nonnegative λ terms weight components. The balance $\lambda_u(k)U_i - \lambda_c(k)C_i$ trades diagnostic refinement against resource use. Training establishes valid diagnosis, introduces grounded information value, then adds cost sensitivity to learn useful acquisition before reducing expenditure. Bounded auxiliary weights prioritize correctness; access restrictions and budgets apply throughout (Appendix B.6).

Group-relative policy updates. The sampling policy π_{old} generates G trajectories from the same initial patient state and budget. Their rewards yield

$$A_i = \frac{R_i^{(k)} - \overline{R}^{(k)}}{\text{std}(R_1^{(k)}, \ldots, R_G^{(k)}) + \varepsilon_A}, \tag{6}$$

where $\overline{R}^{(k)}$ is the group mean and $\varepsilon_A > 0$ prevents division by zero. GRPO clips updates and regularizes toward π_{ref} with KL divergence. Only policy-generated calls, explanations, and diagnosis tokens contribute to loss; tool responses are masked following ReTool (Feng et al., 2026). Each update round synchronizes π_{old} with the policy. Inference selects evidence, analyses, and termination without reward labels or the assessor (Appendix B.7).

4 Experiments

We evaluate diagnostic performance and evidence acquisition, then examine representation, initialization, and policy-training effects.

4.1 Experimental setup

PsyNexus is split by patient into 1,641 training, 205 validation, and 205 test patients (8:1:1). Training uses disjoint SFT, assessor-calibration, and RL subsets of 500, 200, and 941 patients, with approximately 5,000 SFT demonstrations. Each test patient contributes one index encounter, and subsequent label-review records remain inaccessible. Adaptive methods share initial evidence, admissible sources, and resource limits. Baselines cover static models, feature acquisition, and diagnostic agents (Table 2). Appendices C.2 and C.3 provide implementation details.

PsyNexus results are five-seed means on the same test patients. We report macro-F1, macro-AUROC, Brier score, citation precision (CP), and latency. CP measures factual support against acquired sources. The area under the macro-F1 versus budget curve (AUBC) and normalized cost assess acquisition efficiency (Appendix C.9).

Table 2: Diagnostic performance and inference time on PsyNexus. F1 and AUROC are macro averages (%). *Both* denotes EHR and EEG inputs. Time measures end-to-end latency per patient. Results are averaged over five training seeds.

Method	Input	F1 ↑	AUROC ↑	Brier ↓	Time (s) ↓
Static models (full evidence within the stated modalities)					
EHR-only SFT	EHR	77.0	88.5	0.333	16.0
REVE (El Ouahidi et al., 2025)	EEG	68.2	80.7	0.423	1.8
CBraMod (Wang et al., 2025)	EEG	67.3	79.8	0.438	1.7
Static feature fusion	Both	79.1	90.2	0.307	9.0
MUSE (Wu et al., 2024)	Both	80.0	91.0	0.300	10.0
Multimodal SFT	Both	81.2	92.0	0.280	30.0
Adaptive systems (shared budget)					
EDDI (Ma et al., 2019)	Both	75.0	86.7	0.371	8.0
ACO (Valancius et al., 2024)	Both	76.3	88.0	0.351	9.2
LA-CDM (Bani-Harouni et al., 2026)	Both	79.0	90.4	0.318	17.0
CDEG (Dai et al., 2026)	Both	80.5	91.7	0.295	18.0
PRISM-Base (SFT)	Both	77.5	89.2	0.335	16.0
PRISM	Both	86.3	93.9	0.240	12.0

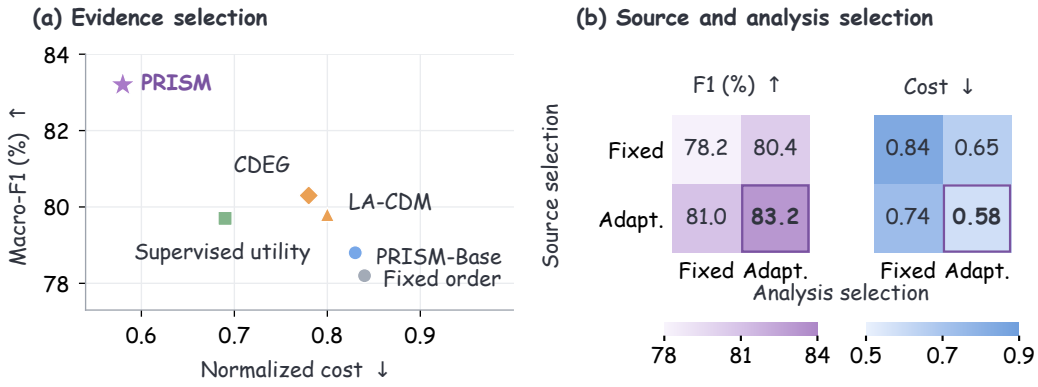


Figure 4: Evidence acquisition under an independent frozen reader. (a) Macro-F1 versus normalized cost. (b) Fixed or adaptive (Adapt.) source and analysis selection. Outlines mark joint adaptation. Values are five-seed means. Full comparisons appear in Figures S4 and S5.

4.2 Main Results

Diagnostic performance. PRISM improves macro-F1 from 81.2% to 86.3% over full-evidence multimodal SFT while reducing latency from 30.0 to 12.0 seconds (Table 2). Relative to LA-CDM (Bani-Harouni et al., 2026), macro-F1 increases from 79.0% to 86.3% and Brier error decreases from 0.318 to 0.240. Higher macro-F1 with lower latency supports resource-efficient diagnosis. The improvement over PRISM-Base supports policy optimization after SFT.

Evidence acquisition. Under a common frozen reader independent of the utility assessor, PRISM-selected evidence achieves 83.2% macro-F1, compared with 79.8% for LA-CDM (Bani-Harouni et al., 2026) and 80.3% for CDEG (Dai et al., 2026) (Figure 4a). Because the reader receives only acquired sources, this comparison separates evidence selection from the policy's generated diagnosis. Gains persist across three readers (Appendix C.6). Joint source and analysis adaptation outperforms either alone (83.2% versus 81.0% and 80.4%) at lower cost (Figure 4b). This supports coordinating both decisions within each patient state.

Evidence-grounded reporting. PRISM achieves 96.0% CP versus 91.0% for LA-CDM, indicating improved factual support alongside diagnostic accuracy. CP measures correspondence to acquired sources rather than clinical completeness or causal faithfulness (Appendix C.10).

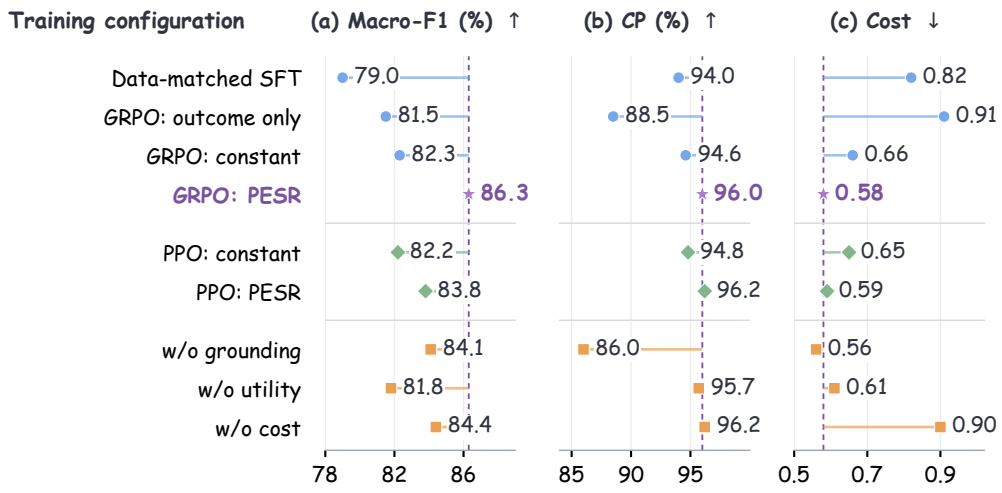


Figure 5: Training and reward comparisons. CP is citation precision, and cost is normalized. Constants match cumulative coefficient exposure. Ablations use GRPO. Dashed lines mark PESR with GRPO. Values are five-seed means. Full results appear in Figures S6 and S7.

4.3 Analysis

Representation and initialization. Figure 6a shows that, with evidence access fixed, adding explicit EEG measurements to clinical evidence and learned representations improves macro-F1 from 79.5% to 81.2% on PsyNexus. The combination also improves REVE (El Ouahidi et al., 2025) probes from 71.5% to 74.0% on TDBRAIN (van Dijk et al., 2022), with consistent gains for CBraMod (Wang et al., 2025). These comparisons support representation complementarity across datasets and encoders (Appendices C.4 and C.11). On matched partial-evidence states, complete SFT raises tool-call validity from 82.0% to 99.3% and numerical fidelity from 74.0% to 98.2% over diagnosis-only supervision, establishing executable and numerically faithful interaction before RL (Appendix C.5). These matched states isolate the contribution of supervised initialization to evidence interpretation and executable tool interaction.

Training strategy and reward components. Figure 5 compares data-matched continued SFT, outcome-only GRPO (Shao et al., 2024), constant rewards, and PESR. Outcome-only GRPO improves macro-F1 over continued SFT but lowers CP from 94.0% to 88.5% and increases cost from 0.82 to 0.91. Optimizing diagnostic correctness alone therefore does not ensure grounded, efficient evidence use. Against exposure-matched constants, PESR improves macro-F1 from 82.3% to 86.3% and reduces cost from 0.66 to 0.58, supporting progressive scheduling beyond cumulative reward exposure. Removing utility reduces F1 to 81.8%, removing grounding lowers CP to 86.0%, and removing cost increases consumption to 0.90. These complementary effects support learning informative, grounded acquisition before optimizing resource expenditure (Appendix C.8).

Optimizer and backbone comparisons. With PPO (Schulman et al., 2017), PESR raises F1 from 82.2% to 83.8% and reduces cost from 0.65 to 0.59 relative to exposure-matched constants (Figure 5). Under GRPO, gains persist across three Qwen3.5 sizes and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) (Figure 6b). Qwen-9B with PESR exceeds 27B with constant rewards (86.3% versus 85.0%), while 27B achieves the highest F1 under PESR. These comparisons support complementary contributions from alignment and model capacity (Appendix C.12).

Qualitative case study. Figure 7 illustrates the distinction between diagnostic correctness and factual support. PRISM-Base overlooks earlier elevated-mood episodes, whereas outcome-only GRPO predicts BD but misreports an EEG value and asserts an unsupported treatment benefit. PRISM revisits the clinical history, reproduces the measurement accurately, and grounds its diagnosis in the longitudinal course, treating the neural observation as complementary evidence.

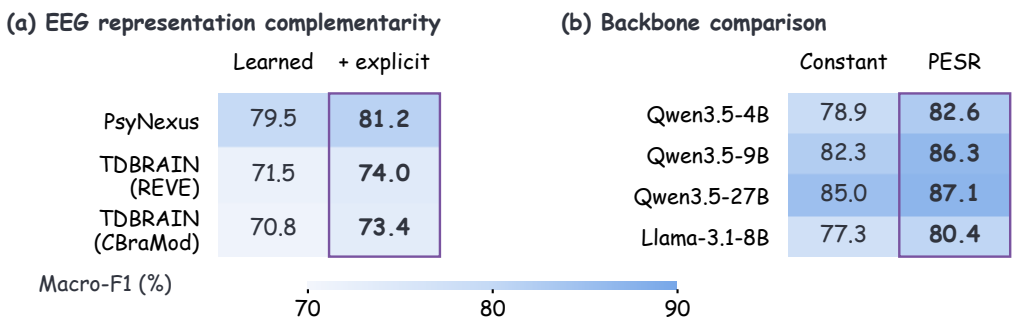


Figure 6: Representation and backbone comparisons. (a) Learned EEG features with and without power and connectivity. PsyNexus includes EHR. (b) PESR versus exposure-matched constant GRPO. Llama-3.1-8B uses the Instruct checkpoint. Macro-F1 (%) comparisons are within rows.

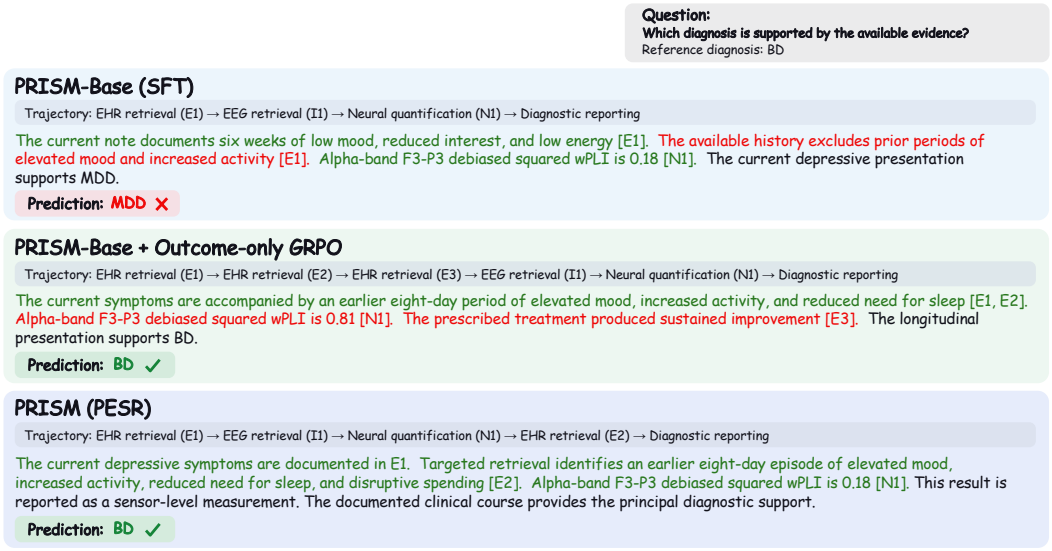


Figure 7: Patient-level evidence acquisition and reporting. Green text marks supported claims, and red text marks unsupported claims or numerical errors. E1–E3 denote EHR sources, I1 an EEG interval, and N1 a neural measurement. The reference diagnosis is used only for evaluation.

Hyperparameter sensitivity. Across tested schedules and individual reward-weight perturbations, macro-F1 ranges from 85.4% to 86.5%. Grounding and cost weights adjust factual support and resource use with limited changes in aggregate accuracy (Appendix C.13).

5 Conclusion

In this work, we present PRISM, an LLM agent framework for patient-specific multimodal psychiatric diagnosis through joint source and analysis selection. PRISM-Base integrates clinical and neural evidence, while PESR progressively aligns evidence acquisition and reasoning with conditional diagnostic value and resource cost. This design enables the agent to adapt information seeking to the evolving patient state and ground diagnostic reports in acquired evidence. Retrospective evaluation on PsyNexus demonstrates a 5.1% absolute macro-F1 improvement and a 60% reduction in inference time relative to full-evidence multimodal SFT. Independent-reader and ablation analyses support joint selection and progressive alignment. More broadly, PRISM offers a methodological foundation for active multimodal agents that jointly learn evidence acquisition and analysis under uncertainty and resource constraints.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments, assist with the interpretation of experimental results, and support translation and language refinement. We have not used generative AI tools to generate synthetic datasets, formulate mathematical claims, provide or write mathematical proofs, or implement the proposed method, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to assist with manuscript drafting and editing, improve readability and organization, and provide feedback on the presentation of figures and experimental results. We have reviewed all AI-assisted work. All technical claims, experimental results, numerical values, and cited literature were independently checked and verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Daman Arora and Andrea Zanette. Training Language Models to Reason Efficiently. In *Advances in Neural Information Processing Systems*, volume 38, pp. 60770–60808. Curran Associates, Inc., 2025. doi: 10.52202/085713-2032.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *International Conference on Learning Representations*, pp. 9112–9141, 2024.

David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Nassir Navab, and Matthias Keicher. Language Agents for Hypothesis-driven Clinical Decision Making with Reinforcement Learning. In *International Conference on Learning Representations*, pp. 156240–156258, 2026.

Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, Qinglin Zhao, Zhenyu Liu, Zhijun Yao, Minqiang Yang, Hong Peng, Jing Zhu, Xiaowei Zhang, Guoping Gao, Fang Zheng, Rui Li, Zhihua Guo, Rong Ma, Jing Yang, Lan Zhang, Xiping Hu, Yumin Li, and Bin Hu. A multi-modal open dataset for mental-disorder analysis. *Scientific Data*, 9(1):178, 2022. doi: 10.1038/s41597-022-01211-x.

Xiwei Dai, Zijie Meng, Zhiting Fan, Yixuan Tang, Guanyu Jiang, Ziru Niu, and Zuozhu Liu. CDEG: Learning Decision-Critical Evidence for Long-Horizon Diagnostic Agents. *arXiv preprint arXiv:2608.22899*, 2026. Version 2.

Hongxin Ding, Baixiang Zhang, Yue Fang, Weibin Liao, Xinke Jiang, Jinyang Zhang, Yinghao Zhu, Zheng Li, Liantao Ma, Junfeng Zhao, and Yasha Wang. ProMed: Shapley Information Gain Guided Reinforcement Learning for Proactive Medical LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32481–32515. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1500.

Yuexi Du, Jinglu Wang, Shujie LIU, Nicha C Dvornek, and Yan Lu. CARE: Towards Clinical Accountability in Multi-Modal Medical Reasoning with an Evidence-Grounded Agentic Framework. In *International Conference on Learning Representations*, pp. 10666–10700, 2026.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A Foundation Model for EEG - Adapting to Any Setup with Large-Scale Pretraining on 25,000 Subjects. In *Advances in Neural Information Processing Systems*, volume 38, pp. 22541–22577. Curran Associates, Inc., 2025. doi: 10.52202/085713-0760.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. ReTool: Reinforcement Learning for Strategic Tool Use in LLMs. In *International Conference on Learning Representations*, pp. 37909–37926, 2026.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep Adaptive Design: Amortizing Sequential Bayesian Experimental Design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3384–3395, 2021.

Tilmann Gneiting and Adrian E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622, 2024. doi: 10.1038/s41591-024-03097-1.

M. D. Hettwer, S. Larivière, B. Y. Park, O. A. van den Heuvel, L. Schmaal, O. A. Andreassen, C. R. K. Ching, M. Hoogman, J. Buitelaar, D. van Rooij, D. J. Veltman, D. J. Stein, B. Franke, T. G. M. van Erp, ENIGMA ADHD Working Group, ENIGMA Autism Working Group, ENIGMA Bipolar Disorder Working Group, ENIGMA Major Depression Working Group, ENIGMA OCD Working Group, ENIGMA Schizophrenia Working Group, N. Jahanshad, P. M. Thompson, S. I. Thomopoulos, R. A. I. Bethlehem, B. C. Bernhardt, S. B. Eickhoff, and S. L. Valk. Coordinated cortical thickness alterations across six neurodevelopmental and psychiatric disorders. *Nature Communications*, 13(1):6851, 2022. doi: 10.1038/s41467-022-34367-6.

Joerg F. Hipp, David J. Hawellek, Maurizio Corbetta, Markus Siegel, and Andreas K. Engel. Large-scale cortical correlation structure of spontaneous oscillatory activity. *Nature Neuroscience*, 15 (6):884–890, 2012. doi: 10.1038/nn.3101.

Jinpeng Hu, Ao Wang, Qianqian Xie, Zhuo Li, Hui Ma, and Dan Guo. AgentMental: An Interactive Multi-Agent Framework for Explainable and Adaptive Mental Health Assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31050–31058, 2026. doi: 10.1609/aaai.v40i37.40365.

Wei-Bang Jiang, Liming Zhao, and Bao-liang Lu. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In *International Conference on Learning Representations*, pp. 16405–16426, 2024.

Wei-Bang Jiang, Yansen Wang, Bao-liang Lu, and Dongsheng Li. NeuroLM: A Universal Multi-task Foundation Model for Bridging the Gap between Language and EEG Signals. In *International Conference on Learning Representations*, pp. 55436–55457, 2025.

Fali Li, Guangying Wang, Sarah Genon, Simon B. Eickhoff, Runyang He, Chanlin Yi, Debo Dong, Dezhong Yao, Lin Jiang, Wei Wu, and Peng Xu. Mapping neurophysiological and molecular profiles of heterogeneity and homogeneity in schizophrenia-bipolar disorder. *Science Advances*, 11(46):eadz0389, 2025. doi: 10.1126/sciadv.adz0389.

Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Scott Yih. RA-DIT: Retrieval-Augmented Dual Instruction Tuning. In *International Conference on Learning Representations*, pp. 19138–19162, 2024.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4234–4243, 2019.

Elzbieta Olejarczyk and Wojciech Jernajczyk. Graph-based analysis of brain connectivity in schizophrenia. *PLOS ONE*, 12(11):e0188629, 2017. doi: 10.1371/journal.pone.0188629.

Su Mi Park, Boram Jeong, Da Young Oh, Chi-Hyun Choi, Hee Yeon Jung, Jun-Young Lee, Donghwan Lee, and Jung-Seok Choi. Identification of Major Psychiatric Disorders From Resting-State Electroencephalography Using a Machine Learning Approach. *Frontiers in Psychiatry*, 12:707581, 2021. doi: 10.3389/fpsyt.2021.707581.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents. Official model release, February 2026.

Khaled Saab, Chunjong Park, Tim Strother, Jan Freyberg, David G. T. Barrett, Yong Cheng, Wei-Hung Weng, David Stutz, Nenad Tomasev, Anil Palepu, Valentin Liévin, Yash Sharma, Roma Ruparel, Abdullah Ahmed, Elahe Vedadi, Kimberly Kanada, Cian Hughes, Yun Liu, Geoff Brown, Yang Gao, Sean Li, S. Sara Mahdavi, James Manyika, Katherine Chou, Yossi Matias, Avinatan Hassidim, Dale R. Webster, Joëlle Barral, S. M. Ali Eslami, Pushmeet Kohli, Adam Rodman, Vivek Natarajan, Mike Schaekermann, Tao Tu, Alan Karthikesalingam, and Ryutaro Tanno. Advancing conversational diagnostic AI with multimodal reasoning. *Nature Medicine*, 32(5):1726–1736, 2026. doi: 10.1038/s41591-026-04371-0.

Dean Salisbury, Dylan Seebold, and Brian Coffman. EEG: First Episode Psychosis vs. Control Resting Task 2. OpenNeuro dataset ds003947, version 1.0.1, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017. doi: 10.48550/arXiv.1707.06347.

Ashlea Segal, Linden Parkes, Kevin Aquino, Seyed Mostafa Kia, Thomas Wolfers, Barbara Franke, Martine Hoogman, Christian F. Beckmann, Lars T. Westlye, Ole A. Andreassen, Andrew Zalesky, Ben J. Harrison, Christopher G. Davey, Carles Soriano-Mas, Narcís Cardoner, Jeggan Tiego, Murat Yücel, Leah Braganza, Chao Suo, Michael Berk, Sue Cotton, Mark A. Bellgrove, Andre F. Marquand, and Alex Fornito. Regional, circuit and network heterogeneity of brain abnormalities in psychiatric disorders. *Nature Neuroscience*, 26(9):1613–1629, 2023. doi: 10.1038/s41593-023-01404-6.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256, 2025. doi: 10.1038/s41586-025-09529-3.

Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. MOTOR: A Time-to-Event Foundation Model For Structured Medical Records. In *International Conference on Learning Representations*, pp. 41038–41077, 2024.

I. Tahir, A. Planat-Chrétien, A. Bertrand, M. Linder, C. Dondé, R. Sosnik, and M. Polosan. Multimodal EEG-fNIRS classification as a clinical tool for bipolar disorder diagnosis. *Translational Psychiatry*, 16(1):177, 2026. doi: 10.1038/s41398-026-03858-1.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, 2025. doi: 10.1038/s41586-025-08866-7.

Michael Valancius, Maxwell Lennon, and Junier Oliva. Acquisition Conditioned Oracle for Nongreedy Active Feature Acquisition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48957–48975, 2024.

Hanneke van Dijk, Guido van Wingen, Damiaan Denys, Sebastian Olbrich, Rosalinde van Ruth, and Martijn Arns. The two decades brainclinics research archive for insights in neurophysiology (TDBRAIN) database. *Scientific Data*, 9(1):333, 2022. doi: 10.1038/s41597-022-01409-z.

Martin Vinck, Marijn van Wingerden, Thilo Womelsdorf, Pascal Fries, and Cyriel M. A. Pennartz. The pairwise phase consistency: A bias-free measure of rhythmic neuronal synchronization. *NeuroImage*, 51(1):112–122, 2010. doi: 10.1016/j.neuroimage.2010.01.073.

Martin Vinck, Robert Oostenveld, Marijn van Wingerden, Francesco Battaglia, and Cyriel M. A. Pennartz. An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. *NeuroImage*, 55(4):1548–1565, 2011. doi: 10.1016/j.neuroimage.2011.01.055.

Dingdong Wang, Shujie Liu, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen Meng. Emotion-Thinker: Prosody-Aware Reinforcement Learning for Explainable Speech Emotion Reasoning. In *International Conference on Learning Representations*, pp. 153708–153733, 2026a.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding. In *International Conference on Learning Representations*, pp. 75310–75346, 2025.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. MedAgent-Pro: Towards Evidence-based Multi-modal Medical Diagnosis via Reasoning Agentic Workflow. In *International Conference on Learning Representations*, pp. 114360–114389, 2026b.

Wei Wu, Yu Zhang, Jing Jiang, Molly V. Lucas, Gregory A. Fonzo, Camarin E. Rolle, Crystal Cooper, Cherise Chin-Fatt, Noralie Krepel, Carena A. Cornelssen, Rachael Wright, Russell T. Toll, Hersh M. Trivedi, Karen Monuszko, Trevor L. Caudle, Kamron Sarhadi, Manish K. Jha, Joseph M. Trombello, Thilo Deckersbach, Phil Adams, Patrick J. McGrath, Myrna M. Weissman, Maurizio Fava, Diego A. Pizzagalli, Martijn Arns, Madhukar H. Trivedi, and Amit Etkin. An electroencephalographic signature predicts antidepressant response in major depression. *Nature Biotechnology*, 38(4):439–447, 2020. doi: 10.1038/s41587-019-0397-3.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Michael Nalls, Jimeng Sun, and Faraz Faghri. Multimodal Patient Representation Learning with Missing Modalities and Labels. In *International Conference on Learning Representations*, pp. 55538–55557, 2024.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Jiaqi Liu, Honglin Guo, Yajie Yang, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, Wei He, Yiwen Ding, Guanyu Li, Zehui Chen, Zhengyin Du, Xuesong Yao, Yufei Xu, Jiecao Chen, Tao Gui, Zuxuan Wu, Qi Zhang, Xuanjing Huang, and Yu-Gang Jiang. AgentGym-RL: An Open-Source Framework to Train LLM Agents for Long-Horizon Decision Making via Multi-Turn RL. In *International Conference on Learning Representations*, pp. 12685–12727, 2026.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Y Zou, and Huaxiu Yao. MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. In *International Conference on Learning Representations*, pp. 66188–66217, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*, 2023.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. CSBrain: A Cross-scale Spatiotemporal Brain Foundation Model for EEG Decoding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 87150–87195. Curran Associates, Inc., 2025. doi: 10.52202/085713-2915.

APPENDIX

Anonymous code repository
https://anonymous.4open.science/r/PRISM-1B77/

A Extended Related Work **15**

A.1 Psychiatric evidence and multimodal representations 15
A.2 Clinical agents and evidence grounding 15
A.3 Sequential design and resource-aware policy alignment 15

B Method Details **16**

B.1 Additional cohort statistics 16
B.2 EHR content and coverage 17
B.3 Patient episodes and access boundaries 18
B.4 Neural measurements and evidence views 18
B.5 Policy initialization and diagnostic outputs 20
B.6 Information value and progressive policy alignment 21
B.7 GRPO objective and training procedure 22
B.8 Matched clinical and neural utility analysis 23
B.9 Execution and resource accounting 23

C Experimental Protocol and Additional Results **24**

C.1 Evaluation structure 24
C.2 Cohort and patient partitions 24
C.3 Baseline definitions and adaptations 24
C.4 Clinical and neural representations 25
C.5 Supervised initialization 25
C.6 Evidence selection across diagnostic readers 27
C.7 Source selection and neural analysis 28
C.8 Progressive alignment and resource allocation 28
C.9 Metric definitions and aggregation 29
C.10 Factual support and interpretation boundaries 30
C.11 EEG representation evaluation on TDBRAIN 30
C.12 Backbone scale and family 31
C.13 Sensitivity to stage boundaries and reward weights 31

A Extended Related Work

This appendix extends Section 2 by examining the representation, interaction, and optimization mechanisms relevant to PRISM. Table 3 summarizes the corresponding evidence types and decision objectives.

A.1 Psychiatric evidence and multimodal representations

Cross-disorder cortical patterns (Hettwer et al., 2022) and heterogeneous EEG connectivity (Li et al., 2025) motivate integrating clinical and neural evidence. Related applications include EEG-based antidepressant response prediction (Wu et al., 2020) and bipolar classification using EEG with functional near-infrared spectroscopy (Tahir et al., 2026). These findings motivate contextual interpretation of neural observations but do not establish universal EEG markers for differential diagnosis among MDD, BD, and SCZ.

Representation learning provides complementary approaches to modeling these observations. LaBraM (Jiang et al., 2024) learns EEG tokens through masked pretraining. CBraMod (Wang et al., 2025) and CSBrain (Zhou et al., 2025) model spatiotemporal structure, while REVE (El Ouahidi et al., 2025) accommodates recording configurations through electrode and time encodings. NeuroLM (Jiang et al., 2025) integrates EEG tokens with language-model instruction tuning. For clinical evidence, MOTOR (Steinberg et al., 2024) learns longitudinal representations, and MUSE (Wu et al., 2024) addresses missing modalities and labels. PRISM builds on multimodal representation to examine patient-specific selection of evidence sources and neural analyses.

A.2 Clinical agents and evidence grounding

Beyond representing observations, diagnostic agents determine how additional evidence is acquired. ReAct (Yao et al., 2023) interleaves reasoning, actions, and observations, while clinical evaluations (Hager et al., 2024) demonstrate the importance of assessing information acquisition beyond examination-style accuracy. AMIE (Tu et al., 2025) develops diagnostic dialogue, and multimodal AMIE (Saab et al., 2026) adds state-aware requests for images and clinical documents. MedAgent-Pro (Wang et al., 2026b) coordinates diagnostic tools, and CARE (Du et al., 2026) evaluates visual evidence and answer consistency. AgentMental (Hu et al., 2026) applies adaptive questioning and memory to depression interviews.

LA-CDM (Bani-Harouni et al., 2026) selects existing textual test reports and cyclically trains hypothesis prediction, confidence calibration, and clinical decisions. PRISM additionally formulates waveform analysis as a policy action. Our LA-CDM adaptation shares the admissible evidence and operations, so empirical comparisons concern learning under common access (Appendix C.3).

Reliable use of acquired evidence additionally requires grounding. Self-RAG (Asai et al., 2024), RA-DIT (Lin et al., 2024), and MMed-RAG (Xia et al., 2025) improve retrieval and contextual evidence use through reflection, dual fine-tuning, or preference training. PRISM distinguishes general medical knowledge from patient-specific observations by linking clinical claims to acquired EHR or EEG evidence throughout the diagnostic trajectory. This requirement motivates evaluating evidence grounding independently of diagnostic correctness.

A.3 Sequential design and resource-aware policy alignment

EDDI (Ma et al., 2019) selects features by expected information gain, while ACO (Valancius et al., 2024) evaluates feature subsets through expected prediction loss and acquisition cost. In medical interaction, ProMed (Ding et al., 2026) rewards questions using contextual Shapley information gain, and CDEG (Dai et al., 2026) uses counterfactual evidence tests to guide acquisition and reappraisal. Deep Adaptive Design (Foster et al., 2021) learns sequential experimental-design policies conditioned on previous observations. For reasoning efficiency, Arora & Zanette (2025) regularize the length of correct responses during reinforcement learning, providing a tunable tradeoff between answer accuracy and token cost. For sequential optimization, EmotionThinker (Wang et al., 2026a) progressively introduces reasoning rewards during GRPO training, while AgentGym-RL (Xi et al., 2026) expands interaction horizons.

Table 3: Evidence acquisition and reasoning in the original methods. PsyNexus baseline adaptations are described in Appendix C.3.

Work	Evidence or action	Selection or supervision
EDDI	Feature acquisition	Expected information gain
ACO	Feature subsets	Expected loss and query cost
Multimodal AMIE	Dialogue and multimodal requests	State-aware dialogue
LA-CDM	Hypotheses and diagnostic tests	Calibration, accuracy, and cost
ProMed	Medical questions	Shapley information gain
CDEG	Acquisition and reappraisal	Counterfactual evidence tests
Arora & Zanette	Response generation	Correctness and length regularization
ReTool	Tool interaction	Outcome-based policy updates
EmotionThinker	Speech emotion reasoning	Progressive reasoning rewards
PRISM (ours)	Clinical acquisition and EEG analysis	Conditional value and resource cost

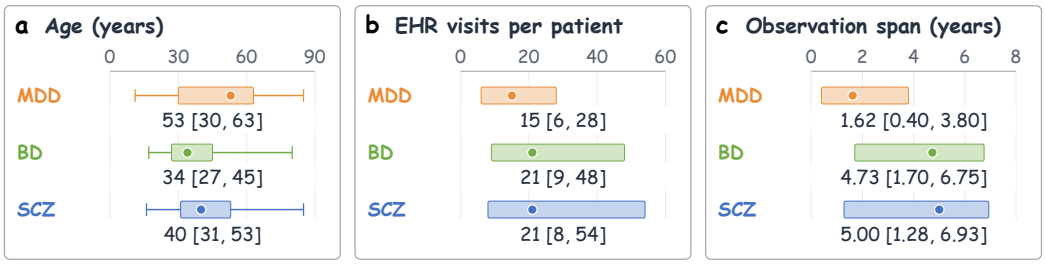


Figure S1: Additional PsyNexus cohort statistics by diagnosis. (a) Age. (b) EHR visits per patient. (c) Observation span. Dots mark medians and boxes span the first to third quartiles (Q1 and Q3), with values shown as median [Q1, Q3]. Age whiskers show the minimum and maximum.

GRPO (Shao et al., 2024) optimizes policies using group-relative reward advantages, while ReTool (Feng et al., 2026) extends outcome-based policy training to multi-turn tool interaction. PESR uses GRPO to align joint source and analysis selection with conditional evidence value and resource cost. A frozen assessor separates the utility criterion from the evolving policy, while progressive rewards introduce cost sensitivity after informative acquisition develops. Appendix B.6 formalizes this objective and schedule.

B Method Details

B.1 Additional cohort statistics

Figure S1 complements Figure 2 with diagnosis-specific distributions of age, EHR encounter count, and observation span. Across PsyNexus, median age is 43 years, with an interquartile range of 29 to 58 years. The median encounter count is 18, with an interquartile range of 8 to 38, and the median observation span is 2.80 years, with an interquartile range of 0.79 to 6.21 years. These distributions characterize the observed cohort without implying a uniform follow-up schedule.

Dataset comparison. In Table 1, FEP Resting EEG abbreviates the descriptive name of OpenNeuro ds003947 (version 1.0.1), officially titled *EEG: First Episode Psychosis vs. Control Resting Task 2*. Participant counts include controls and unlabelled cases where present. Multi. denotes multiple diagnoses or referral indications, with resource-specific diagnostic coverage.

Evidence availability. Column definitions follow Table 1. A ✓ requires corresponding participant-level data for at least some individuals, whereas ✗ denotes absence from the verified release inventory or field schema. Tests excludes research-task scores and raw physiological signals. Recruitment criteria, aggregate statistics, repeated EEG sessions, and recording dates alone

do not establish a treatment archive, longitudinal EHR, or encounter linkage. Availability may vary across participants.

Release verification. We examined official file inventories and participant fields where accessible. MODMA (Cai et al., 2022) provides 53 resting-state EEG files and a participant spreadsheet containing PHQ-9, CTQ-SF, LES, SSRS, GAD-7, and PSQI scores. The BRMH author release (Park et al., 2021) contains a 945-row table of EEG features, demographics, diagnostic labels, and IQ, without source medical records. FEP Resting EEG (Salisbury et al., 2022) includes eight assessment tables and medication doses at recording time expressed as chlorpromazine equivalents. RepOD (Olejarczyk & Jernajczyk, 2017) contains 28 EEG files without participant-level clinical assessment tables. For TDBRAIN, verification includes the release file inventory and participant fields, together with the original data descriptor and repository documentation. Appendix B.2 describes PsyNexus evidence and coverage.

TDBRAIN: release scope and evidence availability. Table 1 uses the 1,274-participant cohort described by van Dijk et al. (2022), whose counts and schema cannot be assumed identical to subsequent releases. The cohort includes healthy and unlabelled participants alongside psychiatric referral indications. Referral indications are distinct from clinician-confirmed diagnoses, and UNKNOWN does not identify healthy controls. The original Table 1 counts EEG sessions by indication and permits multiple indications per participant. Consequently, it does not provide mutually exclusive patient counts for a three-disorder benchmark.

The released BDF files contain 26 scalp EEG channels sampled at 500 Hz under two-minute eyes-open and eyes-closed conditions, with additional EOG, ECG, and EMG channels. The original Table 2 reports BDI scores before and after treatment for 176 participants, ADHD-RS for 102, Y-BOCS for 48, and item-level NEO-FFI data for 1,018 sessions. These source-reported counts do not imply unrestricted availability for every participant, as the release reserves 30% of known MDD, ADHD, and OCD diagnostic and prognostic categories for blinded validation. The Tx checkmark denotes repetitive transcranial magnetic stimulation (rTMS) and neurofeedback protocols and treatment outcomes. Medication use was explicitly not tracked systematically.

The release inventory and participant fields contain no narrative clinical notes, laboratory results, clinical examination reports, longitudinal EHR, or EEG-to-EHR encounter mappings. Notes, Tests, Long. EHR, and Link are therefore marked absent under the criteria above. EEG session identifiers organize recordings without linking them to clinical encounters.

B.2 EHR CONTENT AND COVERAGE

PsyNexus comprises longitudinal outpatient and inpatient EHR, integrating structured records with narrative clinical documentation. The categories below summarize the source archive. Coverage varies across patients, encounters, and clinical settings, and the presence of a category does not imply that every patient has all corresponding assessments, examinations, or treatments.

Demographics and encounters. Records include age, sex, marital status, occupation, encounter type, department, and visit dates. Inpatient records additionally describe admission and discharge dates, admission route, discharge disposition, and length of stay. Longitudinal summaries include age at first visit, encounter counts, and first and last visit dates.

History and clinical assessment. Outpatient, emergency, and admission notes document presenting complaints, the history of the current illness, and past medical, personal, and family histories. Some records additionally describe surgical, trauma, and allergy histories. Clinical assessments include physical, mental-status, and neurological examinations, with findings on cognition, affect, volition, behavior, and neurological function.

Diagnoses and clinical course. Diagnostic records contain names, codes, categories, and dates, including admission and discharge diagnoses. Progress notes, ward-round notes, and clinical discussion records document changes during hospitalization. Discharge summaries describe the condition at admission, treatment course, response, discharge condition, and recommendations. These

source-archive fields remain subject to the target-field masking and temporal restrictions specified in Appendix B.3.

Orders and treatment. Medication orders specify drug formulations, quantities, doses, units, administration frequency and route, intended duration, and order start and stop times. These records describe prescribed therapy without establishing administration or adherence. Treatment records additionally document planned and completed session counts, execution times, and departments. Recorded procedures include modified electroconvulsive therapy, transcranial magnetic stimulation, and acupuncture.

Laboratory and examination reports. Laboratory records include test names, results, units, reference ranges, specimen types, and timestamps. They cover hematology, urinalysis, biochemistry, coagulation, and endocrine testing. Microbiology records include organisms, colony counts, and antimicrobial susceptibility results. Examination reports contain findings, conclusions, and dates for EEG, electrocardiography, computed tomography, magnetic resonance imaging, and ultrasonography. Report availability does not establish access to the corresponding raw images or signals.

Vital signs and nursing observations. Recorded items include temperature, pulse, respiratory rate, blood pressure, oxygen saturation, glucose, height, weight, body mass index, pain, and sleep. Observations retain values, units, and recording times where available.

Psychometric assessments. Records include scale names, assessment items and results, order and completion times, total or subscale scores, and assessment comments. Instruments assess general symptoms (SCL-90), sleep quality (PSQI), depressive symptoms (SDS), anxiety symptoms (SAS), social functioning (SDSS), mood-disorder screening (MDQ), and suicidal behavior (SBQ-R). Some inpatient documents additionally record positive- and negative-symptom assessments (SAPS and SANS), insight and treatment-attitude assessments (ITAQ), and risk assessments. Instrument availability does not imply complete item-level data or repeated measurements for every patient.

B.3 Patient episodes and access boundaries

Each diagnostic episode specifies a patient, an index encounter, an input cutoff, initial evidence, an admissible archive, and a reference diagnosis. The task is retrospective identification of a longitudinally reviewed disorder from evidence available at that encounter. Reference-label review considers diagnostic records and supporting clinical documentation across follow-up, whereas policy inputs are restricted to records available by the cutoff. The adjudication record is required to identify source diagnoses, their dates, diagnostic revisions, and the rule for resolving disagreements. Reference labels cannot be inferred from model predictions or EEG features. Cases with unresolved diagnoses require an explicit eligibility decision and documentation in the split manifest.

The earliest eligible encounter is selected according to a rule fixed before partition-level evaluation. The task does not reproduce the diagnosis originally recorded at the index encounter. Diagnosis codes, diagnostic statements, and assessment conclusions that directly reveal the target are masked in evidence and metadata. Event time and record availability are stored separately. Records with unknown availability times remain inaccessible, without assigning inferred dates.

To preserve this access boundary, source discovery exposes only permitted modality, existence, and recording-time metadata. Content-bearing titles, summaries, derived features, and retrieval excerpts require acquisition, and candidate generation cannot inspect cached EEG values before the corresponding operation. Requests for inaccessible content are blocked without revealing hidden information. Preprocessing, reference distributions, dictionaries, and retrieval stores are fitted on training patients. All sources, windows, branches, and encounters associated with a patient remain within one partition. The primary evaluation uses one index encounter per patient.

B.4 Neural measurements and evidence views

Analysis intervals. Each interval j belongs to one recording and recording condition. Measurements are computed from nonoverlapping artifact-screened windows and aggregated within the interval, without pooling encounters. The frequency bands are δ (0.5–4.5 Hz), θ (4.5–8.5 Hz), α

(8.5–11.5 Hz), σ (11.5–15.5 Hz), and β (15.5–30 Hz). Each band includes its lower boundary and excludes its upper boundary, except that β includes 30 Hz. Valid estimation requires acquisition and filtering to retain 0.5–1 Hz content, as subsequent feature extraction cannot recover excluded frequencies. The reference, filters, window duration, tapers, minimum valid data, and channel policy are fixed in the analysis configuration.

Quality metadata record valid duration, valid window count, bad channels, artifact fraction, and available recording conditions. Estimates with insufficient support receive validity masks, preserving the distinction between a valid numerical zero and an unavailable value. Connectivity diagonals are excluded from evidence. When an encoder uses interpolated channels, those channels remain explicitly marked and do not contribute purportedly observed connections to the explicit measurement branch.

Coherence and phase consistency. Let $S_{uv}(f)$ be the mean cross-spectrum at frequency f over valid windows, and let $S_{uu}(f), S_{vv}(f)$ be the corresponding power spectra. Coherence magnitude is

$$\mathrm{Coh}_{uv}(f) = \frac{|S_{uv}(f)|}{\sqrt{S_{uu}(f)S_{vv}(f)}}. \tag{7}$$

Let K be the number of valid spectral observations and $z_k(f) = \exp(\mathrm{i}[\varphi_{u,k}(f) - \varphi_{v,k}(f)])$, where $\varphi_{u,k}$ is channel u's phase in observation k and $\mathrm{i}^2 = -1$. PPC (Vinck et al., 2010) is

$$\mathrm{PPC}_{uv}(f) = \frac{|\sum_{k=1}^{K} z_k(f)|^2 - K}{K(K-1)}. \tag{8}$$

The estimate requires $K > 1$, and the spectral implementation specifies how taper contributions are combined within windows. Tapers do not constitute independent patients or clinical observations.

Lag-sensitive phase coupling. Let $a_k(f)$ be the imaginary component of the cross-periodogram for window k. Debiased squared wPLI (Vinck et al., 2011) is

$$\mathrm{dwPLI}^2_{uv}(f) = \frac{(\sum_k a_k(f))^2 - \sum_k a_k(f)^2}{(\sum_k |a_k(f)|)^2 - \sum_k a_k(f)^2}. \tag{9}$$

A zero denominator renders the estimate undefined. Finite-sample negative values of PPC and debiased squared wPLI are retained without interpreting them as negative coupling. Spectral estimates are computed across windows before frequency bins are averaged within each band. Coherence is reported as magnitude, without squaring.

Orthogonalized envelopes. For complex analytic signals $z_u(t), z_v(t)$ in a band, pairwise orthogonalization (Hipp et al., 2012) uses

$$z_{v\perp u}(t) = \mathrm{Im}\left(z_v(t) \frac{z_u(t)^*}{|z_u(t)|} \right), \qquad r_{v|u} = \mathrm{corr}(|z_u|, |z_{v\perp u}|), \tag{10}$$

where t denotes signal time in this equation, $*$ denotes complex conjugation, and corr is Pearson correlation over valid time samples. Symmetric oAEC is obtained by averaging $|r_{v|u}|$ and $|r_{u|v}|$, followed by aggregation over valid windows using the fixed rule. Zero envelopes and constant inputs yield invalid estimates. Amplitude envelopes are used without logarithmic transformation. This estimator reduces particular instantaneous mixing effects but does not establish source localization or causality.

Relative band power is computed as the integral of the channel power spectrum within a band divided by its 0.5–30 Hz integral over the same interval. Denominators, units, and validity masks are retained. Channel-group summaries specify group membership, exclude self-connections, and report the number of valid constituent entries. A group summary and its constituent edges are not treated as independent observations.

Queries and provenance. A measurement query q returns $e_q = (j, \tau, b, r, U, v_q, m_q, \eta_q)$, where τ denotes the interval, b the band, r the estimator, U the channel or channel-pair set, v_q the numerical values, m_q the validity masks, and η_q the quality metadata and estimator settings. Source identifiers link e_q to the waveform and preprocessing version. Cohort-relative comparisons additionally record the reference population fitted on training patients. Without such a reference, observations are reported as measurements without classification as clinically abnormal.

B.5 POLICY INITIALIZATION AND DIAGNOSTIC OUTPUTS

Tool interface and model inputs. Clinical evidence retrieval, EEG segment retrieval, neural feature quantification, EEG representation encoding, and diagnostic reporting correspond to `read_ehr`, `open_eeg`, `measure_eeg`, `encode_eeg`, and `stop`, respectively. Clinical events retain source spans, assertion status, and temporal fields. Clinical records and queried measurements are serialized as text with source, availability-time, and query identifiers. The interface exposes only requested entries or predefined channel-group summaries, with validity masks and quality metadata for unavailable estimates. Neural tokens are pooled to a fixed count and mapped into the language-model embedding space through a two-layer projector with layer normalization. The projector processes retrieved intervals only. Following an encoding operation, projected tokens are incorporated into the evidence context with their interval identifiers.

Patient selection and demonstration coverage. The SFT subset in Appendix C.2 covers MDD, BD, and SCZ, including the minority diagnosis, with variation in clinical course, record completeness, and clinical and EEG evidence availability. Demonstrations span evidence combinations, trajectory lengths, and tool operations. Multiple demonstrations increase interaction coverage without increasing the number of independent patients.

Executable supervision. Under the access rules in Appendix B.3, a sampler selects admissible clinical sources, EEG intervals, and predefined analyses using permitted metadata, varying trajectory length and modality coverage within the resource budget. Measurement and encoding require prior interval retrieval. Sampled calls execute through the RL tool interface, and their partial-evidence prefixes define supervised training states. Demonstration counts refer to complete examples rather than derived prefixes.

Each prefix provides evidence targets containing clinical spans, source identifiers, and numerical fields, tool targets specifying executable calls and arguments, and diagnostic targets containing the reference disease code and source-supported evidence fields. Retained examples require replayable calls, citations resolving to acquired sources, and numerical targets consistent with tool responses. Clinical interpretations require separate evidence verification, with unsupported statements excluded from supervision.

Diagnostic state and outputs. A fixed output schema specifies tool calls, the final diagnosis, class probabilities, and cited evidence. Three tokenizer-verified single-token codes represent MDD, BD, and SCZ, with a shared code mapping and diagnostic prefix across methods. Diagnostic prediction uses canonical acquired evidence, while action selection additionally conditions on class probabilities, access history, and remaining budgets. Each observation is indexed by source and query. Successful acquisition updates the evidence, access history, budget, and diagnostic probabilities without updating model parameters during inference. Probabilities derive from logits restricted to the disease codes rather than generated confidence estimates.

Supervised initialization optimizes the policy adapters, EEG projector, and an auxiliary three-class assessor head. Its loss is

$$\mathcal{L}_{\mathrm{SFT}} = \mathcal{L}_{\mathrm{diag}} + \alpha_{\mathrm{tool}}\mathcal{L}_{\mathrm{tool}} + \sum_k \alpha_k \mathcal{L}_k, \qquad \mathcal{L}_k = \frac{\sum_n m_{n,k}\ell_k(\widehat{\zeta}_{n,k}, \zeta^{\mathrm{ref}}_{n,k})}{\max(1, \sum_n m_{n,k})}, \tag{11}$$

where $\mathcal{L}_{\mathrm{diag}}$ sums the cross-entropy losses for the policy class token and auxiliary assessor head, $\mathcal{L}_{\mathrm{tool}}$ supervises admissible tool schemas, k indexes evidence fields, and n indexes annotation slots. $m_{n,k}$ identifies a verified reference $\zeta^{\mathrm{ref}}_{n,k}$, $\widehat{\zeta}_{n,k}$ is its prediction, ℓ_k is the field-specific loss, and the nonnegative α coefficients weight the supervision terms. Training includes admissible partial-evidence states both with and without EEG tokens.

The SFT checkpoint initializes the trainable policy π_θ, frozen reference π_{ref}, and frozen diagnostic assessor $D_{\bar{\psi}}$. The assessor retains its auxiliary head and undergoes temperature scaling (Guo et al., 2017) on the disjoint calibration subset in Table 4. With network parameters fixed, a single positive temperature T_{cal} minimizes negative log likelihood, averaged within patients before aggregation across patients. Calibration states cover initial evidence, partial clinical evidence, and combined clinical and EEG evidence under a fixed sampling protocol.

T_{cal} remains fixed throughout RL. Policy optimization updates only the adapters, leaving the reference, assessor, encoder, and projector unchanged. The final policy is subsequently calibrated using validation patients and frozen before testing. Its diagnostic probabilities remain distinct from those of the reward assessor.

B.6 Information value and progressive policy alignment

Sequential design perspective. PESR adopts a sequential experimental-design perspective (Foster et al., 2021) for retrieval and analysis of existing records. Each admissible operation determines which observation becomes available for subsequent reasoning. Trajectory-level optimization learns acquisition, analysis, and termination jointly, allowing an intermediate operation to incur cost when the resulting evidence sequence improves the final return. The deployed policy amortizes these decisions without evaluating counterfactual outcomes online.

Alignment objective. The term $\lambda_u(k)U_i - \lambda_c(k)C_i$ in Eq. 5 balances diagnostic refinement against resource expenditure. The weight bounds below preserve correctness as the primary reward.

Factual grounding. Let M_i be the number of distinct atomic evidence claims in response i, and let $v_{i,j} \in [0,1]$ be the independently verified support for claim j. We use

$$R_{\text{ground},i} = \frac{1}{\max(1, M_i)} \sum_{j=1}^{M_i} v_{i,j}. \tag{12}$$

Verification compares source identifiers, temporal and assertion attributes, and numerical fields against the acquired evidence. Duplicate claims do not increase M_i, unsupported claims receive zero credit, and empty responses receive zero grounding reward. Only claims represented in the scored schema are eligible for verification. Narrative interpretations lacking reliable verification do not receive automatic clinical-correctness labels.

Dynamic information value. For operation t in trajectory i, the signed gain is

$$\Delta_{i,t} = F(E_{i,t+1}, y) - F(E_{i,t}, y), \qquad U_i = \frac{1}{2} \sum_{t=0}^{T_i - 1} \Delta_{i,t}. \tag{13}$$

The sum telescopes to Eq. 4, so U_i measures total evidence value under the frozen assessor. The dependence of $\Delta_{i,t}$ on $E_{i,t}$ makes this value dynamic: the same observation can have different utility after different clinical or neural evidence. The reward coefficients follow a shared training schedule, while the evidence state determines conditional value. Identical repeated evidence has zero incremental utility. Signed gains, including zero and negative values, remain recorded. The assessor evaluates tool-provided evidence while excluding policy-generated explanations, and labels enter external reward calculation only.

Brier error is a strictly proper scoring loss (Gneiting & Raftery, 2007), whose expected value is minimized by the true conditional class distribution. Under an accurate conditional model, its expected optimal loss measures quadratic predictive uncertainty. In PESR, the learned assessor provides an approximation, and U_i measures realized, label-referenced probabilistic improvement. Correcting an overconfident error can improve this score while increasing predictive entropy. Thus, diagnostic refinement need not monotonically increase confidence, and U_i is not Shannon information gain. Policy learning uses realized returns to improve expected acquisition behavior without an explicit per-action gain predictor.

Cost and execution validity. Let $c_{i,r}$ be cumulative cost in resource dimension r, $b_{0,r} > 0$ its initial budget, and $w_r \geq 0$ fixed weights summing to one. We use

$$C_i = \sum_{r=1}^{d} w_r \frac{c_{i,r}}{b_{0,r}}, \tag{14}$$

where d is the number of resource dimensions. Clinical access volume, EEG duration, and computation follow the accounting rules in Appendix B.9. Admissible, budget-compliant trajectories satisfy $C_i \in [0,1]$.

In Eq. 5, V_i counts invalid operations or output events, each subject to the fixed penalty λ_v. Missing final diagnoses and invalid response schemas receive zero outcome and grounding credit, while recorded acquisition utility, incurred costs, and execution penalties are retained. Failure handling follows the fixed rules in Appendix B.9.

Progressive alignment schedule. Let $q_k = \frac{k}{K_{\text{train}}}$ be the fraction of completed updates, where K_{train} is the total number of scheduled RL updates, and let $0 < a < b < 1$ be stage boundaries. For $\mathcal{R}(x) = \min(1, \max(0, x))$, we set

$$\begin{aligned} \lambda_g(k) &= \lambda_g^* \mathcal{R}\left(\frac{q_k - a}{b - a}\right), \qquad \lambda_u(k) = \lambda_u^* \mathcal{R}\left(\frac{q_k - a}{b - a}\right), \\ \lambda_c(k) &= \lambda_c^* \mathcal{R}\left(\frac{q_k - b}{1 - b}\right), \end{aligned} \tag{15}$$

where starred coefficients denote final reward weights. Unless varied in Appendix C.13, the reference configuration is $(a, b) = (0.20, 0.60)$ and $(\lambda_g^*, \lambda_u^*, \lambda_c^*) = (0.20, 0.15, 0.10)$. Outcome and execution terms remain active throughout training. For $q_k \leq a$, grounding, utility, and cost coefficients are zero. Grounding and utility coefficients then increase linearly until $q_k = b$, after which they remain fixed while the cost coefficient increases. This ordering first develops supported, informative acquisition and then introduces pressure for economical resource allocation. Appendix C.8 defines the fixed-weight and exposure-matched controls.

For admissible trajectories, $R_{\text{ground}} \in [0, 1]$, $U \in [-1, 1]$, and $C \in [0, 1]$. The final weights satisfy $\lambda_g^* + 2\lambda_u^* + \lambda_c^* < 1$. Consequently, any correct admissible trajectory receives a reward of at least $1 - \lambda_u^* - \lambda_c^*$, whereas an incorrect admissible trajectory receives at most $\lambda_g^* + \lambda_u^*$. The strictly positive difference preserves correctness as the primary reward throughout the schedule, without guaranteeing policy convergence or clinical generalization.

B.7 GRPO OBJECTIVE AND TRAINING PROCEDURE

For each initial state, the sampling policy π_{old} generates G trajectories. It remains fixed during batch collection and optimization, then synchronizes with the updated policy before the next batch. In contrast, the SFT reference π_{ref} remains fixed throughout RL. Let $\mathcal{I}_i$ denote policy-generated token positions in trajectory i, $z_{i,n}$ the generated token, and $h_{i,n}$ its preceding visible context. The policy ratio is

$$r_{i,n}(\theta) = \frac{\pi_\theta(z_{i,n} \mid h_{i,n})}{\pi_{\text{old}}(z_{i,n} \mid h_{i,n})}. \tag{16}$$

Using A_i from Eq. 6, GRPO (Shao et al., 2024) maximizes

$$\begin{aligned} J_{\text{GRPO}}(\theta) &= \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathcal{I}_i|}\sum_{n\in\mathcal{I}_i}\left(L_{i,n}(\theta) - \beta K_{i,n}(\theta)\right)\right], \\ L_{i,n}(\theta) &= \min(r_{i,n}(\theta)A_i, \text{clip}(r_{i,n}(\theta), 1-\epsilon, 1+\epsilon)A_i), \\ K_{i,n}(\theta) &= D_{\text{KL}}(\pi_\theta(\cdot \mid h_{i,n}) \,\|\, \pi_{\text{ref}}(\cdot \mid h_{i,n})), \end{aligned} \tag{17}$$

where ϵ is the clipping parameter, β weights KL regularization, and D_{KL} denotes Kullback-Leibler divergence. The expectation is taken over training patients and sampled trajectories. The trainable, sampling, and reference policies use identical admissible-token masks at constrained positions. The implementation records the exact or sampled KL estimator, and external tool observations are excluded from KL computation.

Tool calls, final class tokens, and policy-generated evidence explanations contribute to the loss. Fixed prompts, EHR excerpts, measurements, encoder outputs, and environment-generated error messages are masked, following the feedback-masking principle in ReTool (Feng et al., 2026), which originally uses PPO. The objective is normalized by the number of policy-generated tokens within each trajectory. Groups with identical rewards have zero reward advantage and remain included in trajectory accounting.

Each update round uses fresh on-policy trajectories. Reusing earlier trajectories as current-policy samples would require a separately specified correction. RL sampling uses fixed diagnosis-stratified

Algorithm 1 PRISM training with PESR and budgeted inference

Require: Patient-level split (1,641 training, 205 validation, 205 test), disjoint training subsets (500 SFT, 200 calibration, 941 RL), tools, budgets, reward schedule

Initialize PRISM-Base and the assessor head using Eq. 11
Calibrate the assessor on the 200-patient subset using patient-weighted temperature scaling
Freeze $D_{\bar{\psi}}$, π_{ref}, the EEG encoder, and the projector
for each RL update round k **do**
    Snapshot $\pi_{\text{old}} \leftarrow \pi_\theta$
    for each sampled RL patient state s_0 **do**
        Sample G tool-use trajectories under identical initial budgets
        Compute outcome, grounding, utility, cost, and violation scores
        Form $R_i^{(k)}$ and A_i using Eqs. 5 and 6
    end for
    Update policy adapters with Eq. 17, masking tool observations
end for
Select and calibrate the final policy on validation patients, then freeze it
for each test patient **do**
    while acquisition budget remains **do**
        Select a legal operation or stop using π_θ
        if stop is selected **then**
            break
        end if
        Execute the operation, update the patient state, and charge actual cost
    end while
    Return diagnosis, class probabilities, and cited evidence within the reserved output budget
end for

probabilities with equal patient weights within each stratum, while preserving the test distribution. Checkpoint selection, reward weights, and decoding settings are determined exclusively using validation patients.

B.8 Matched clinical and neural utility analysis

For clinical operation c and neural operation q that are both feasible from the same initial evidence E, define

$$\Gamma_{c,q} = F(E + c + q, y) - F(E + c, y) - F(E + q, y) + F(E, y). \tag{18}$$

The difference in the unnormalized gain from q after and before c equals $\Gamma_{c,q}$. All comparison branches share the frozen assessor, calibration, evidence serialization, and query parameters. If EEG retrieval is a prerequisite, comparisons begin after the interval has been retrieved. The interaction may be positive, zero, or negative and does not identify a biological causal effect.

This contrast characterizes conditional evidence value under matched observations. It is used for analysis and does not introduce a separate learned selector or an additional reward term in Eq. 5. The training objective does not require every patient to acquire both modalities. Identical final evidence yields identical assessor utility, although acquisition order may affect feasibility and execution cost. Appendix C.6 evaluates acquired evidence with an independent diagnostic reader.

B.9 Execution and resource accounting

Resource accounting records unique EHR tokens, unique EEG seconds, and total computation. Additional diagnostics include LLM input and output tokens and encoder calls. Newly accessed signal duration is the union of retrieved intervals. Repeated acquisition of identical content adds no evidence or unique access volume, but repeated context processing, measurement, and generation remain computational costs. Training uses deterministic operation counts or calibrated serial processing costs, excluding unrelated shared-resource queue delays. End-to-end evaluation includes source discovery, diagnostic-state calculation, policy inference, tool execution, and final output. De-

cleared limits and timeouts enforce resource budgets. A separate reserve for the final diagnostic report is fixed before training trajectories are collected.

Failed operations and unavailable responses add no evidence, while their incurred costs remain recorded. Timeouts preserve previously acquired valid evidence. Reader-based reevaluation retains the original selector's consumption. The environment distinguishes archive exhaustion, policy-selected termination, and budget exhaustion. During deployment, the policy requires neither ground-truth labels nor the frozen reward assessor. Online preprocessing and encoding contribute to inference cost, whereas training-time reward evaluation and reference-model computation are accounted for separately.

C Experimental Protocol and Additional Results

This appendix details patient partitions, baseline adaptations, metric definitions, and complete numerical comparisons. PsyNexus results are averaged over five training seeds evaluated on the same 205 independent test patients.

C.1 Evaluation structure

The evaluation separates overall diagnostic performance (Table 2) from acquisition quality under an independent reader (Figure 4) and the effects of training objectives and reward components (Figure 5). Fixed-evidence comparisons isolate representation effects, while matched partial-evidence states assess supervised initialization. Final-report citation precision additionally evaluates factual support independently of diagnostic correctness.

C.2 Cohort and patient partitions

The 2,051 patients are assigned to training, validation, and test sets of 1,641, 205, and 205 patients, respectively. All records and derived examples from a patient remain in one partition. Each test patient contributes one index episode under the eligibility and input-cutoff rules in Appendix B.3.

Within training, all 500 SFT patients contribute to parameter estimation through approximately 5,000 demonstrations, 200 separate patients calibrate the utility assessor, and the remaining 941 patients support RL (Table 4). These disjoint subsets separate assessor fitting and calibration from policy optimization. Calibration estimates a single temperature with network parameters fixed (Appendix B.5). Validation patients support model selection and final-policy calibration. Baselines may use the combined SFT and RL population of 1,441 patients under their respective objectives. Assessor-calibration patients are excluded from gradient-based training of all compared systems.

All adaptive methods receive demographics and the masked chief complaint as initial evidence, with missing fields retained as unavailable. A shared operation catalogue specifies admissible source requests, EEG intervals, channel groups, and measurement types under the access restrictions in Appendix B.3. Test patients are excluded from retrieval memories, evidence graphs, feature normalization, and hyperparameter selection.

C.3 Baseline definitions and adaptations

Static diagnostic models. EHR-only SFT fine-tunes Qwen3.5-9B (Qwen Team, 2026) using clinical evidence alone. REVE (El Ouahidi et al., 2025) and CBraMod (Wang et al., 2025) use pooled EEG representations with task-specific three-class classification heads. Static feature fusion concatenates clinical embeddings, relative power, predefined connectivity summaries, pooled REVE features, and availability masks. Feature normalization and the fusion head are fitted on training patients.

MUSE (Wu et al., 2024) retains its patient-modality bipartite graph and mutual-consistent contrastive objective. Its adapted inputs comprise clinical embeddings, explicit EEG measurements, and learned EEG features, with missing branches masked. Graph construction uses training patients and excludes test labels and test-time graph fitting. The comparison therefore evaluates adaptation to PsyNexus rather than replication of the original task.

Table 4: Cohort allocation and model configuration. SFT, assessor calibration, and RL use disjoint subsets of the training partition.

Item	Setting
Training, validation, and test patients	1,641, 205, and 205
SFT, calibration, and RL patients	500, 200, and 941
SFT demonstrations	Approximately 5,000
Training seeds	5
Policy backbone	Qwen3.5-9B (Qwen Team, 2026)
EEG encoder	REVE (El Ouahidi et al., 2025) (frozen)
EEG channels, frequency bands, and connectivity measures	19, 5, and 4

Multimodal SFT uses Qwen3.5-9B (Qwen Team, 2026), the same EEG encoder and projector architecture, and supervised diagnostic and source-grounding objectives over all eligible evidence. It shares PRISM's eligible training-patient pool. Inputs exceeding the context limit undergo source-linked chunking and aggregation, with every processing pass included in latency. Full evidence denotes eligible evidence within each model's specified modalities. Accordingly, EEG classifiers exclude EHR, and static models are evaluated without the acquisition-budget constraint imposed on adaptive methods.

Adaptive systems. PRISM-Base uses the SFT interaction policy without RL. EDDI (Ma et al., 2019) combines Partial VAE conditional inference with target information gain, while ACO (Valancius et al., 2024) retains acquisition-conditioned subset evaluation of expected prediction loss and cost. Both methods operate on predefined clinical and neural feature groups mapped to the shared tool interface, with conditional expectations estimated from training-derived distributions. The SFT diagnostic reader evaluates their acquired evidence in Table 2.

LA-CDM (Bani-Harouni et al., 2026) uses hypothesis and decision roles for uncertainty-aware, cost-sensitive acquisition. Both roles use the Qwen3.5-9B backbone family, and all invocations contribute to inference cost. CDEG (Dai et al., 2026) retains trajectory contrasts, counterfactual evidence validation, and graph-guided acquisition, with graphs and memories constructed from training patients. Reappraisal may reconsider acquired evidence but cannot expose inaccessible sources. All adaptive methods share admissible evidence, tool prerequisites, initial states, and resource limits. Independent-reader evaluation additionally controls the diagnostic model applied to selected evidence. The LA-CDM comparison evaluates its PsyNexus adaptation under this shared operation catalogue, including EEG analysis, rather than the published MIMIC-CDM setting. Differences therefore cannot be attributed to exclusive EEG access for PRISM.

C.4 Clinical and neural representations

To isolate representation effects, Figure S2 compares clinical evidence E, relative band power P, connectivity measurements M, and learned EEG representations H. These supervised comparisons fix evidence access, patients, available intervals, backbone, and diagnostic targets. The EEG encoder remains frozen, and the projector is trained when H is included. All configurations distinguish numerical zeros from unavailable measurements.

The combined representation achieves 81.2% macro-F1 and 92.0% macro-AUROC, with a Brier score of 0.280. Adding explicit power and connectivity to clinical evidence increases macro-F1 from 77.0% to 79.7%, and incorporating learned representations further increases it to 81.2%. Conversely, explicit measurements improve $E + H$ from 79.5% to 81.2%. These reciprocal gains indicate complementary diagnostic information in explicit and learned EEG representations. Numerical fidelity exceeds 97% for every configuration containing explicit measurements, indicating that their inclusion preserves numerical consistency.

C.5 Supervised initialization

To separate supervised initialization from acquisition behavior, we compare diagnosis-only supervision, diagnosis with tool supervision, and complete SFT on identical held-out partial-evidence states. The variants share the supervised patient subset and diagnostic targets. Complete SFT additionally

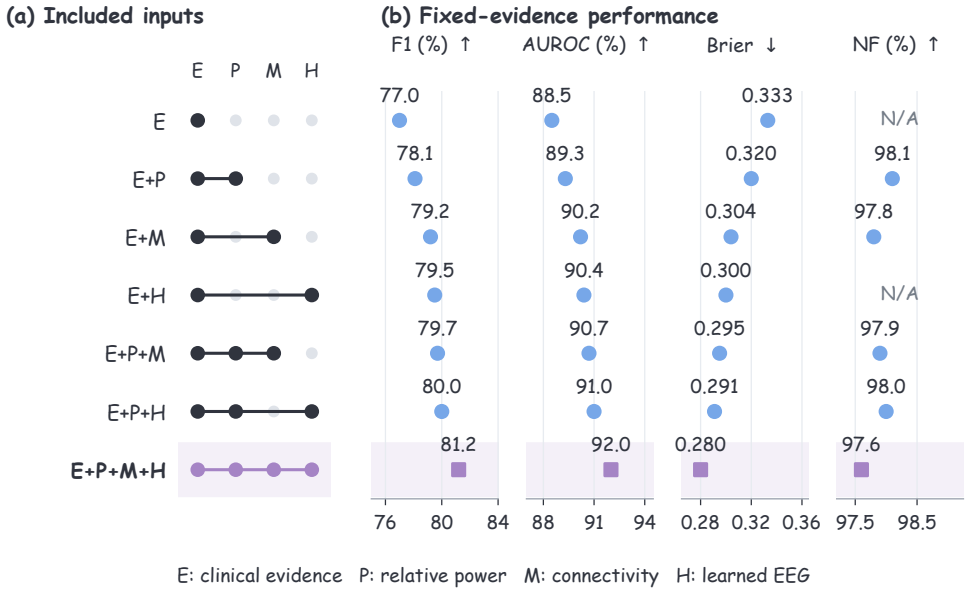


Figure S2: Fixed-evidence representations on PsyNexus. Dark and purple nodes mark included inputs: clinical evidence E, relative power P, connectivity M, and learned EEG features H. Markers show macro-F1, macro-AUROC, Brier score, and numerical fidelity (NF). N/A marks configurations without explicit numerical EEG claims. Values are five-seed means.

supervises verified evidence fields and source references. Holding the available observations fixed isolates evidence interpretation and executable interaction from differences in evidence acquisition.

Figure S3 shows that tool supervision increases the valid-call rate from 82.0% to 98.5%. Complete SFT further achieves 95.0% citation precision, 98.2% numerical fidelity, and 92.0% evidence-slot recall, indicating that evidence supervision improves factual reproduction as well as tool execution. Its 76.0% macro-F1 is evaluated on partial-evidence states, whereas the 77.5% PRISM-Base result in Table 2 concerns complete diagnostic trajectories. The two results therefore characterize distinct evidence conditions.

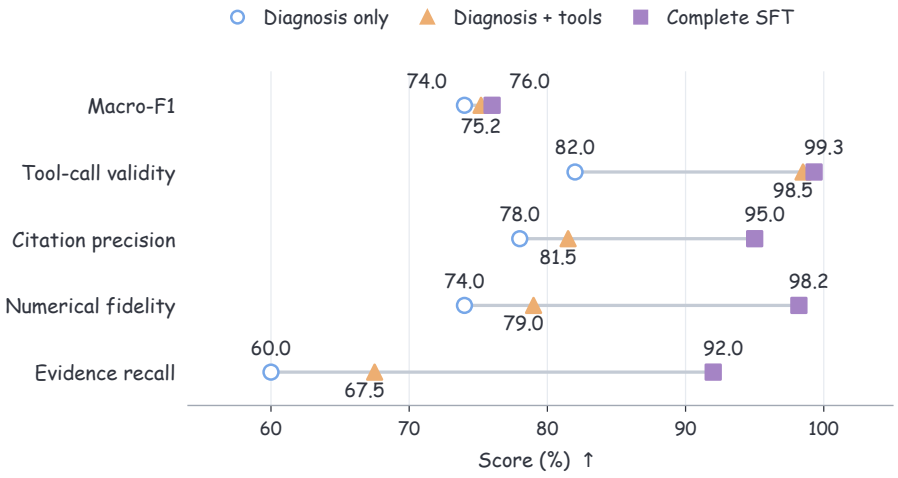


Figure S3: PRISM-Base supervision on matched partial-evidence states. Lines connect the three supervision variants. Tool-call validity measures executable calls. Evidence recall measures recovery of reference evidence slots in the presented state. Scores are five-seed means in percent.

C.6 EVIDENCE SELECTION ACROSS DIAGNOSTIC READERS

To assess whether acquisition benefits persist across diagnostic models, evidence from each selector is evaluated by three frozen readers: the SFT model, the final RL model, and an independently trained reader distinct from the utility assessor. Readers receive acquired source evidence without the selector's generated diagnosis. Acquisition cost remains that of the original trajectory, irrespective of subsequent reader evaluation. AUBC in Figure S4 uses the independent reader at each budget.

The control policies separate patient-specific acquisition from simpler selection rules. Fixed order uses a patient-invariant operation sequence, while random selection samples admissible operations. Metadata-only selection relies on permitted source-availability and recording metadata without accessing unacquired content. Supervised utility selection predicts acquisition value from supervised targets. PRISM-Base and PRISM use their SFT and PESR-trained acquisition policies, respectively.

PRISM-selected evidence achieves macro-F1 scores of 80.9%, 86.3%, and 83.2% under the SFT, RL, and independent readers, respectively. LA-CDM (Bani-Harouni et al., 2026) achieves 79.0%, 81.1%, and 79.8%, while CDEG (Dai et al., 2026) achieves 80.5%, 82.0%, and 80.3%. PRISM's improvement over LA-CDM persists when the diagnostic reader is held fixed, supporting an acquisition benefit beyond differences between their diagnostic models. Under the independent reader, PRISM also improves AUBC from 0.767 to 0.804 at lower normalized cost (0.58 versus 0.80). These comparisons evaluate complete selectors, while the following factorial ablation isolates source and analysis adaptation within PRISM.

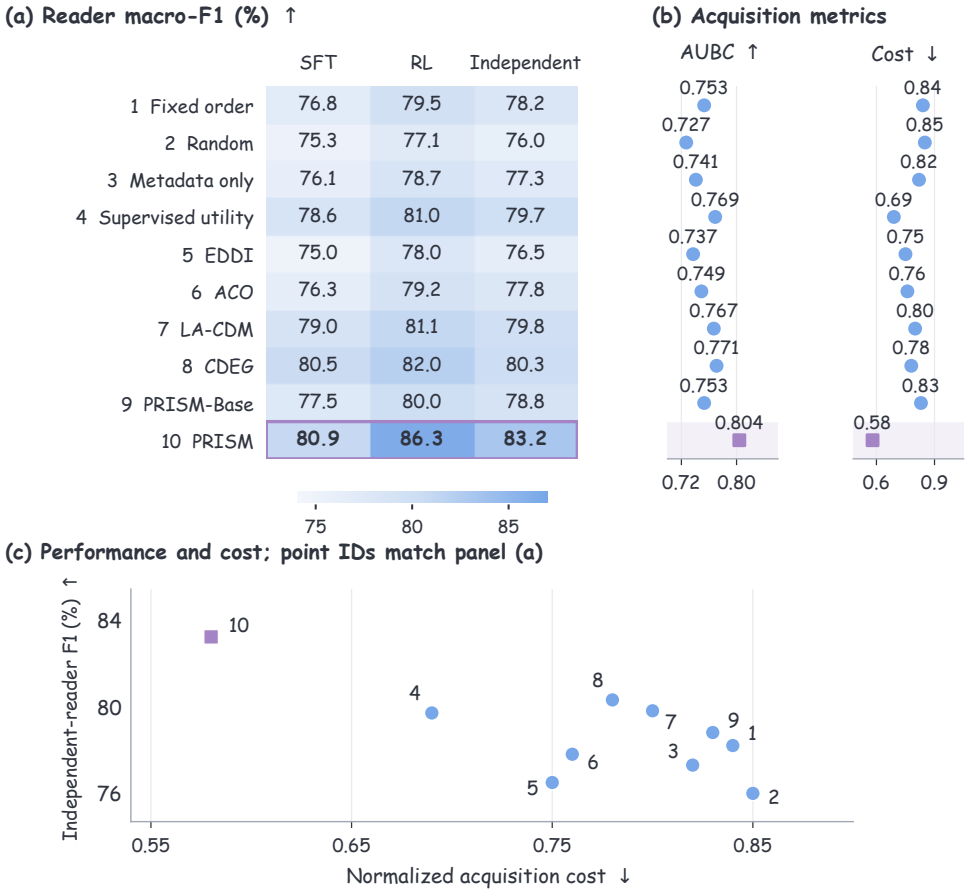


Figure S4: Evidence acquisition under frozen diagnostic readers. (a) Macro-F1 (%) under SFT, RL, and independently trained readers. (b) AUBC integrates independent-reader macro-F1 over budget. Cost is normalized trajectory consumption. (c) Independent-reader F1 versus cost, with point IDs matching panel (a). Values are five-seed means. Controls follow Appendix C.6.

C.7 Source selection and neural analysis

To distinguish the two decision types, we independently enable patient-specific source and analysis selection in a 2×2 comparison. Fixed source selection uses a global ordering, while fixed analysis applies a shared neural-analysis configuration. Adaptive alternatives condition the corresponding decision on the patient state. Each single-adaptation variant holds the other decision fixed, with admissible operations and resource constraints shared across variants. All four configurations use the independent frozen reader from Figure S4.

Figure S5 shows that both forms of adaptation improve performance over fixed selection. Adaptive source selection increases macro-F1 from 78.2% to 81.0%, while adaptive analysis increases it to 80.4%. Joint adaptation achieves 83.2% macro-F1 and 0.804 AUBC while reducing normalized cost from 0.84 to 0.58. These findings support jointly allocating resources to informative sources and analyses. The joint policy invokes the EEG encoder for 44% of patients, compared with 100% under fixed analysis, consistent with patient-specific selection of neural representations.

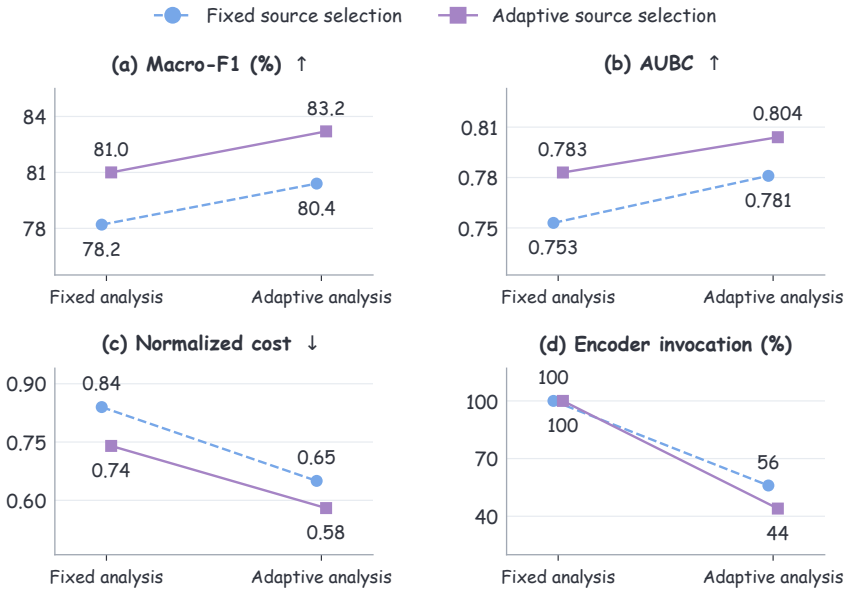


Figure S5: Source and analysis selection. Each line fixes source selection while varying analysis selection. (a) Macro-F1. (b) Independent-reader area under the budget curve (AUBC). (c) Normalized cost. (d) Percentage of episodes with at least one EEG encoding call. Values are five-seed means under the reader and cost definitions in Figure S4.

C.8 Progressive alignment and resource allocation

To isolate optimization effects, all training controls share the SFT initialization and eligible training patients. Data-matched continued SFT controls for additional supervised training on the RL population, while outcome-only GRPO (Shao et al., 2024) isolates diagnostic-correctness optimization. Fixed final weights activate all reward terms from the first update. Exposure-matched constant weights equate each coefficient's cumulative exposure with that of progressive scheduling. Component ablations separately test information-value feedback, cost sensitivity, and factual reliability by removing utility, cost, or grounding while retaining the remaining terms.

Figure S6 reports all training variants. PESR achieves higher macro-F1 and AUBC than both fixed-weight controls, supporting the progression from information-value learning to cost-sensitive allocation. Removing utility reduces macro-F1 to 81.8%, indicating the role of value feedback. Omitting grounding reduces macro-F1 from 86.3% to 84.1% and CP by 10.0 percentage points, whereas omitting cost reduces macro-F1 to 84.4% and increases normalized consumption from 0.58 to 0.90.

Although omitting cost yields slightly higher AUBC (0.810 versus 0.807), full PESR achieves higher final macro-F1 with lower resource consumption. These results support an empirical trade-off between diagnostic value and resource expenditure under factual reliability. They do not imply simultaneous optimality across every performance and cost metric.

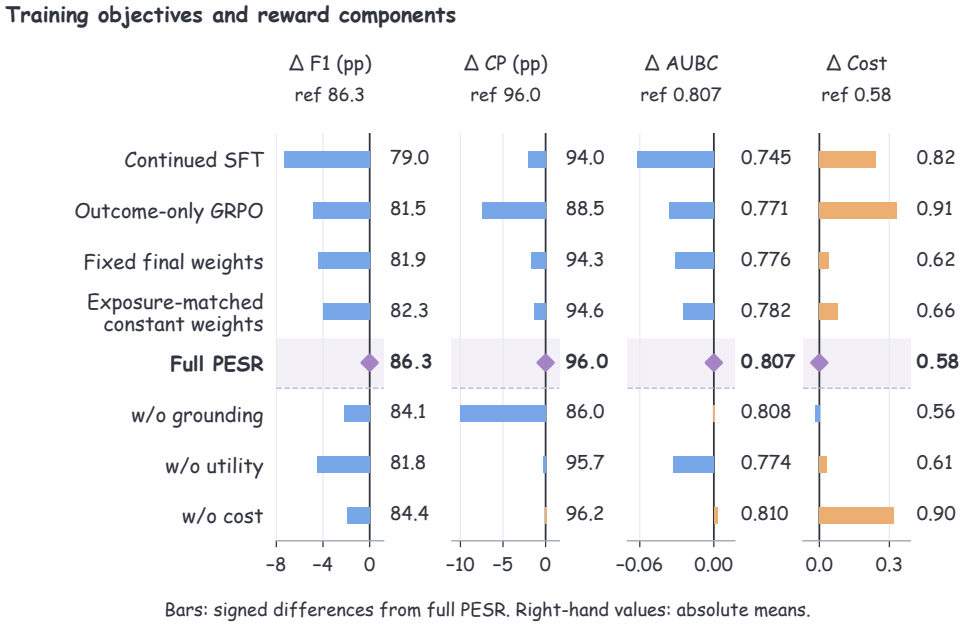


Figure S6: Training controls and GRPO reward ablations. Bars show signed differences from full PESR, and right-hand labels report absolute five-seed means. F1 is macro-F1, CP citation precision, AUBC area under the budget curve, and pp percentage points. Higher F1, CP, and AUBC and lower cost are preferred. AUBC uses each variant's diagnostic model (Appendix C.8).

Optimizer and reward strategies. Figure S7 crosses PPO (Schulman et al., 2017) and GRPO with outcome-only, exposure-matched constant, and progressive PESR rewards. The six configurations use matched initialization and tool budgets. Under PPO, PESR improves over the constant-reward control by 1.6 F1 points and 1.4 CP points while reducing cost by 0.06. This supports progressive rewards beyond GRPO, although the displayed means do not establish statistical equivalence between optimizers.

C.9 Metric definitions and aggregation

The following definitions and aggregation procedures apply to the PsyNexus evaluation.

Diagnostic performance. Macro-F1 averages class-specific F1 scores for MDD, BD, and SCZ, while macro-AUROC averages one-vs-rest AUROCs. For patient i, let p_i denote the predicted three-class distribution and $\mathbf{e}_{y_i}$ the one-hot reference label. The multiclass Brier score is $N^{-1}\sum_{i=1}^{N}\|p_i - \mathbf{e}_{y_i}\|_2^2$, where N is the number of test patients. This unhalved squared error ranges from 0 to 2. F1, AUROC, and evidence-quality rates use a 0–100 scale. Prediction metrics are computed independently for each training seed before averaging across five seeds.

Evidence interpretation. Citation precision measures the proportion of cited factual claims supported by acquired sources at the diagnostic cutoff, using the verification criteria in Appendix B.6. Empty reports receive no credit. Numerical fidelity evaluates reported values, units, frequency bands, estimators, channels, intervals, and source identifiers against recorded tool responses, using serialization precision to set the tolerance. This metric applies to explicit numerical EEG claims, with N/A assigned to configurations lacking such claims. Valid-call rate measures the proportion

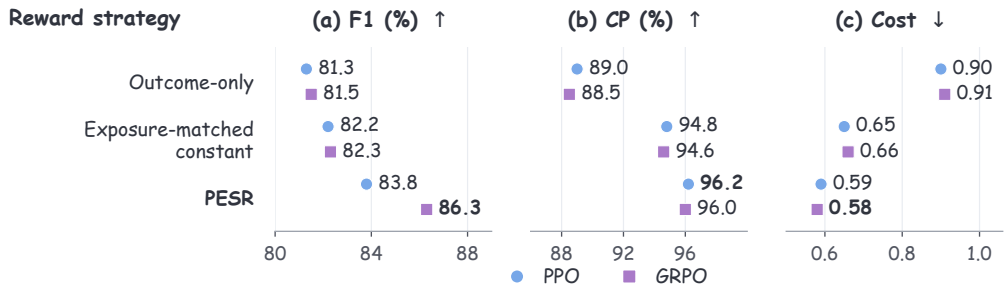


Figure S7: PPO and GRPO under matched patient splits, initialization, and tool budgets. (a) Macro-F1. (b) Citation precision (CP). (c) Normalized cost. Constant rewards match cumulative coefficient exposure. Values are five-seed means.

of executable calls satisfying the tool schema and operation prerequisites. Evidence-slot recall in Figure S3 quantifies recovery of reference slots within the presented partial-evidence state. Its scope excludes complete-archive coverage and final-report completeness after adaptive acquisition.

Budgets and resource use. Fractions $\beta_l \in \{0, 0.1, 0.25, 0.5, 0.75, 1\}$ scale fixed limits on incremental clinical access, EEG duration, and computation. At zero budget, only initial evidence and the reserved diagnostic output are available. Let f_l denote macro-F1 on the 0–1 scale at budget fraction β_l. The area under the budget curve is

$$\mathrm{AUBC} = \frac{1}{\beta_L - \beta_0} \sum_{l=1}^{L} (\beta_l - \beta_{l-1}) \frac{f_l + f_{l-1}}{2}, \tag{19}$$

where L is the number of adjacent budget intervals. All budget conditions use the same test patients, and AUBC is calculated for each seed before averaging. The independent-reader AUBC of 0.804 in Figure S4 and policy-specific AUBC of 0.807 in Figure S6 use different diagnostic readers and therefore quantify distinct evaluation settings.

Normalized cost follows Eq. 14, while Table 2 reports end-to-end latency in seconds per patient. Both use the accounting rules in Appendix B.9. The patient is the independent evaluation unit, with repeated states and trajectories aggregated within patients before population-level summaries. All five training seeds use the same patient partition.

C.10 Factual support and interpretation boundaries

Figure S8 compares citation precision across report-generating systems from the overall evaluation and outcome-only GRPO. PRISM achieves the highest CP among these methods. Together with the grounding ablation in Figure S6, the comparison shows that factual support requires separate assessment even when diagnostic performance remains high.

Scope of the evaluation. The experiments assess retrospective three-disorder classification on PsyNexus and do not establish prospective clinical effectiveness or generalization of the complete agent to external institutions. Citation precision measures source support without establishing clinically sufficient explanations. Sensor-level connectivity and relative power do not identify anatomical sources or causal disease mechanisms. Independent-reader conclusions remain conditional on the evaluated readers and evidence interface. Resource measures quantify processing of archived records rather than human cognitive effort, patient burden, or the number of clinical procedures. The results do not establish globally optimal information acquisition or minimum resource use.

C.11 EEG representation evaluation on TDBRAIN

Figure S9 compares fixed-evidence EEG representations on TDBRAIN (van Dijk et al., 2022). Relative power P and connectivity M constitute explicit features, while H_R and H_C denote frozen representations from REVE (El Ouahidi et al., 2025) and CBraMod (Wang et al., 2025), respectively. Adding explicit features improves macro-F1 from 71.5% to 74.0% for REVE and from 70.8% to

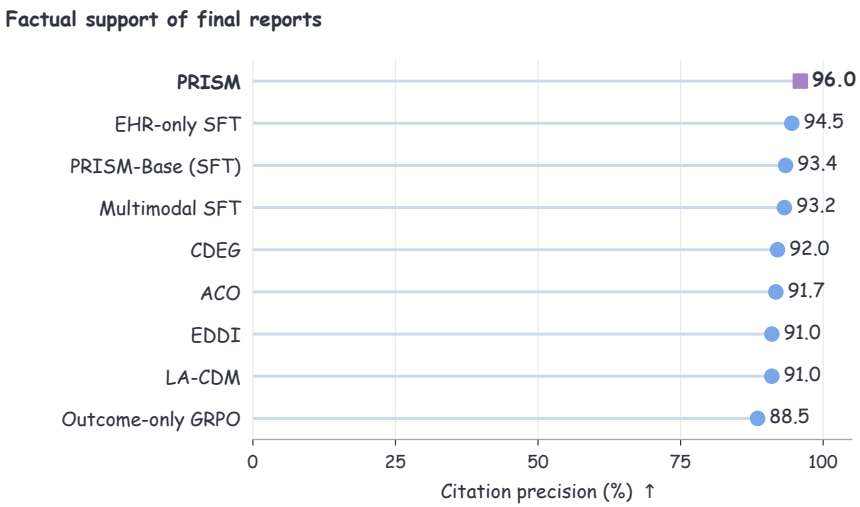


Figure S8: Final-report citation precision on PsyNexus (%), ranked by five-seed mean. Citation precision assesses source correspondence rather than completeness of the clinical explanation.

73.4% for CBraMod, with corresponding improvements in balanced accuracy, AUROC, and Brier score. Within the REVE combinations, adding connectivity to $P + H_R$ further increases macro-F1 from 72.0% to 74.0%. These comparisons support complementary information in explicit and learned neural representations. They evaluate the neural representation component rather than the complete PRISM acquisition policy.

C.12 Backbone scale and family

Figure S10 tests Qwen3.5 at 4B, 9B, and 27B and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). PESR outperforms all four controls on every backbone. Qwen reaches 82.6%, 86.3%, and 87.1% F1, exceeding exposure-matched constants by 3.7, 4.0, and 2.1 points: absolute performance rises with size, whereas the scheduling gain is nonmonotonic. Llama improves from 77.3% to 80.4% (+3.1 points), supporting benefits beyond Qwen. Qwen-9B with PESR also exceeds 27B with constant rewards (86.3% versus 85.0%), showing that alignment offsets part of the observed scale gap in this comparison.

C.13 Sensitivity to stage boundaries and reward weights

We use Qwen3.5-9B, the main experiment's patient partition and tool budget, and five-seed means on 205 test patients. The reference configuration is $(a, b) = (0.20, 0.60)$ and $(\lambda_g^*, \lambda_u^*, \lambda_c^*) = (0.20, 0.15, 0.10)$. It balances the three metrics rather than optimizing them simultaneously.

Stage boundaries. Figure S11 varies $a \in \{0.10, 0.20, 0.30\}$ and $b \in \{0.50, 0.60, 0.70\}$ at fixed final weights. Boundaries denote fractions of completed training updates. Macro-F1 spans 85.5–86.3%, CP 95.0–96.2%, and cost 0.54–0.65. Later b increases cost at each a. This sweep changes both schedule timing and cumulative coefficient exposure.

Final reward weights. We apply multipliers $m \in \{0.5, 0.75, 1, 1.25, 1.5\}$ to one coefficient at a time (Figure S12), holding other weights and (a, b) fixed. Across the 13 configurations, macro-F1 spans 85.4–86.5%. Increasing λ_g^* from 0.10 to 0.30 raises CP from 93.8% to 96.9%. Increasing λ_c^* from 0.05 to 0.15 reduces cost from 0.76 to 0.48 while F1 decreases from 86.5% to 85.8%. Utility perturbations yield F1 of 85.4–86.4%. These sweeps do not establish robustness to joint perturbations or statistical equivalence. All tested weights satisfy $\lambda_g^* + 2\lambda_u^* + \lambda_c^* < 1$.

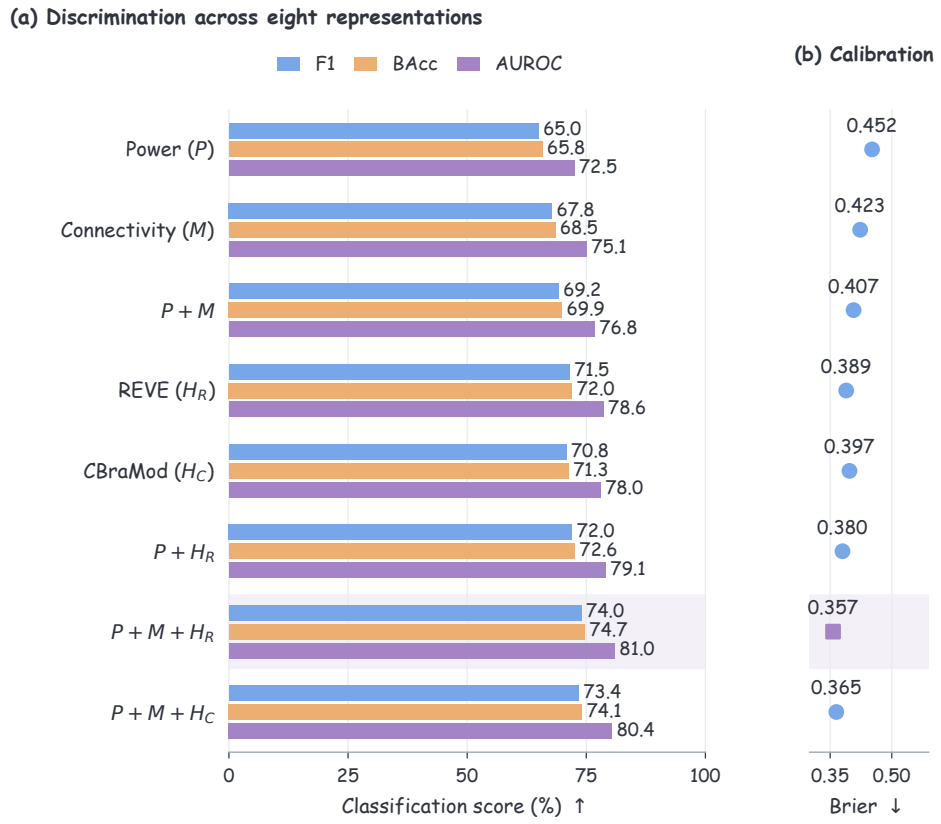


Figure S9: Fixed-evidence classification probes on TDBRAIN. (a) Macro-F1, balanced accuracy (BAcc), and AUROC in percent, ordered from top to bottom within each representation. (b) Brier score. P is relative power, M is connectivity, and H_R and H_C are frozen REVE and CBraMod representations.

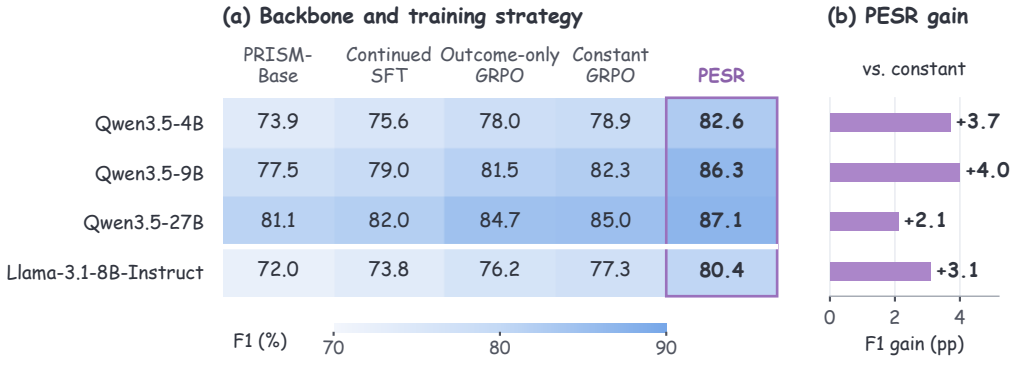


Figure S10: Backbone scale and family comparison on PsyNexus. Panel (a) reports macro-F1 (%), and panel (b) shows PESR gains over exposure-matched constant rewards in percentage points (pp). All configurations share the patient split and tool budget. Values are five-seed means.

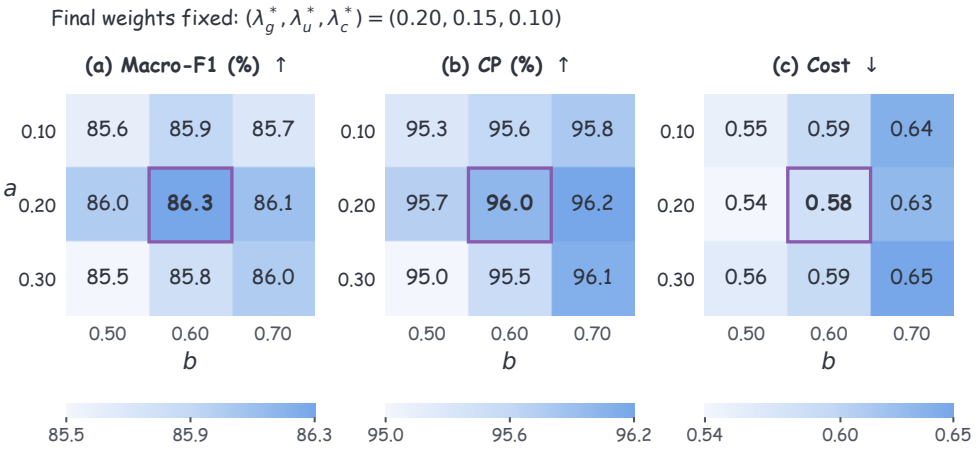


Figure S11: Stage-boundary sensitivity at fixed final weights $(\lambda_g^*, \lambda_u^*, \lambda_c^*) = (0.20, 0.15, 0.10)$. Stage transitions a and b are fractions of total updates. Panels show macro-F1 (%), citation precision (CP, %), and normalized cost. Outlines mark $(a, b) = (0.20, 0.60)$. Values are five-seed means.

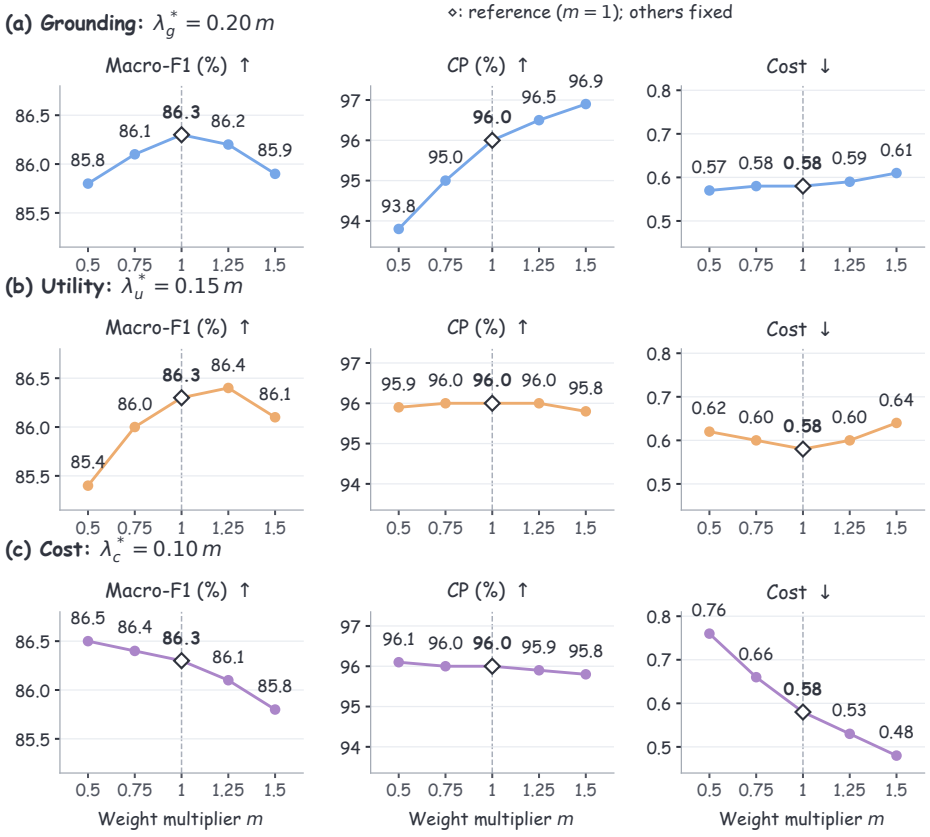


Figure S12: Individual reward-weight sensitivity at $(a, b) = (0.20, 0.60)$. Each row scales its indicated coefficient by m, holding the others fixed. CP is citation precision, and cost is normalized. Diamonds mark $m = 1$. Columns share metric-specific vertical scales. Values are five-seed means, and lines connect tested settings.