Blackbox to Glassbox: Interpretable Clinical Trial Matching via LLM Probing

HS-SOFT-371

1 INTRODUCTION

Clinical trial matching pairs patients to eligibility criteria. Trials that cannot recruit are **delayed or abandoned**. LLMs can automate this but are **unreliable and opaque**.

RESEARCH QUESTIONS

1. Can a probed open-source (8B) model reliably verify patient eligibility?
2. Can we mechanistically explain which criteria drove the decision?

Approach: Formulate as a verification problem. Use **linear probes** for decisions and **mechanistic circuit extraction** for explainability.

2 METHODS

1 Dataset

Five trials from ClinicalTrials.gov (Iron, Osteoporosis, Kidney, Coagulation, SLE). 50 train / 100 test rows per trial. Paired-row design: each patient appears with correct and incorrect assessment.

2 Model

Qwen3-8B (8B params, 36 layers, 32 heads/layer, d=4096). Activations extracted at 36 layers × 1 token position = 36 layer probes per trial, plus 36 × 32 head probes.

3 PROBLEM FORMULATION

Problem Formulation

Definition 1 (Clinical Trial Eligibility Assessment Verification)**.** *Let $\mathcal{T}$ be a clinical trial with inclusion criteria $\mathcal{I} = \{I_1, \ldots, I_p\}$ and exclusion criteria $\mathcal{E} = \{E_1, \ldots, E_q\}$. Let $\mathbf{r}$ be a patient record containing demographic, clinical, and laboratory data. Let $\mathcal{M}$ be a large language model that produces an eligibility assessment $a \in \{\texttt{eligible}, \texttt{excluded}\}$ given $(\mathbf{r}, \mathcal{T})$.*

*The **correctness verification task** is to determine whether the assessment a is consistent with the evidence in $\mathbf{r}$ with respect to the criteria $(\mathcal{I}, \mathcal{E})$:*

$$y = \begin{cases} 1 \text{ (correct)} & \text{if } a = \texttt{eligible} \wedge \bigwedge_{k=1}^{p} sat(\mathbf{r}, I_k) \wedge \bigwedge_{k=1}^{q} \neg sat(\mathbf{r}, E_k) \\ 1 \text{ (correct)} & \text{if } a = \texttt{excluded} \wedge \left(\bigvee_{k=1}^{p} \neg sat(\mathbf{r}, I_k) \vee \bigvee_{k=1}^{q} sat(\mathbf{r}, E_k)\right) \\ 0 \text{ (incorrect)} & \text{otherwise} \end{cases} \quad (1)$$

where $sat(\mathbf{r}, C)$ denotes that patient record $\mathbf{r}$ satisfies criterion C.

Formulation by the author.

4 DATASET & DESIGN

Dataset Construction

Clinical Trials

Trial ID	Domain	Key Criteria
NCT00378469	Iron	Male, 18–40, BMI 18–25
NCT00252421	Osteo.	Female, ≥50, BMD 0 to −2.0
NCT00556491	Kidney	>18, cardiac surg., GFR 15–90
NCT00618579	Coag.	Male, 18–45, FVIII ≥50%
NCT00371501	SLE	>18, ≥4 ACR, ≥2 CV risk

50 train / 100 test rows per trial.

Criteria sourced from ClinicalTrials.gov. Table by the author.

Paired-Row Design

Each patient appears in **two rows**:

$$\mathbf{s}_i^+ = (\mathbf{r}_i \| \mathcal{T}_k,\ a_{\text{correct}}) \qquad y^+ = 1$$
$$\mathbf{s}_i^- = (\mathbf{r}_i \| \mathcal{T}_k,\ a_{\text{incorrect}}) \qquad y^- = 0$$

- $\mathbf{r}_i$ = patient record
- $\mathcal{T}_k$ = trial criteria
- a_{correct} = ground-truth assessment
- $a_{\text{incorrect}}$ = opposite
- y = label (1 = correct, 0 = incorrect)

Probe learns correctness, not patient features.

Balanced: equal correct/incorrect. Formulation by the author.

5 ALGORITHMS

Algorithms

Algorithm 1: Probe Selection

Input: $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$
Output: Best probe f^*, layer l^*, token t^*

```
/* Step 1: Train                                   */
for l = 1 to L do
    for t = 1 to T do
        Extract a_{l,t} from residual stream ;
        Train probe f_{l,t} on activations ;
/* Step 2: Evaluate                                */
for each probe f_{l,t} do
    Compute F1(f_{l,t}, D_val) ;
/* Step 3: Select best                             */
(l*, t*) ← arg max_{l,t} F1 ;
return f_{l*,t*}, l*, t*
```

Algorithm 2: Circuit Discovery

Input: 1,152 attention heads, prompt
Output: Decision circuit (small set of heads)

```
/* Step 1: Score each head                         */
for each head do
    sig ← attn to relevant tokens ;
    noise ← attn to other tokens ;
    sel ← sig / (sig + noise) ;
    score ← sel × sig ;
/* Step 2: Three role-specific scans               */
Scan A: relevant = criteria ;
Scan B: relevant = patient meds ;
Scan C: relevant = meds at output ;
Circuit ← top heads per scan ;
/* Step 3: Validate                                */
Validate on 5 patients ;
```

Algorithms by the author.

6 RESULTS: PROBE ACCURACY

89.9%	96.8%	84.0%	70.4%
LINEAR PROBE	GPT-4O	CLAUDE HAIKU	GPT-4O-MINI

Results

Trial Diversity vs. Accuracy

Configuration	Trials	Accuracy
Multi-trial (TREC)	500	50.0%
5-trial combined	5	72.0%
3-trial combined	3	87.5%
Individual (mean)	1	**89.9%**

Fewer trials → better generalization.

Table by the author.

Per-Trial Probe Results (Unseen Patients)

Trial ID	Domain	Layer	Acc.
NCT00378469	Iron	24	95.5%
NCT00618579	Coag.	22	93.0%
NCT00371501	SLE	22	88.0%
NCT00252421	Osteo.	21	87.0%
NCT00556491	Kidney	24	86.0%
Mean	—		**89.9%**
Cross-trial transfer	—		55.0%

100 test rows/trial. Table by the author.

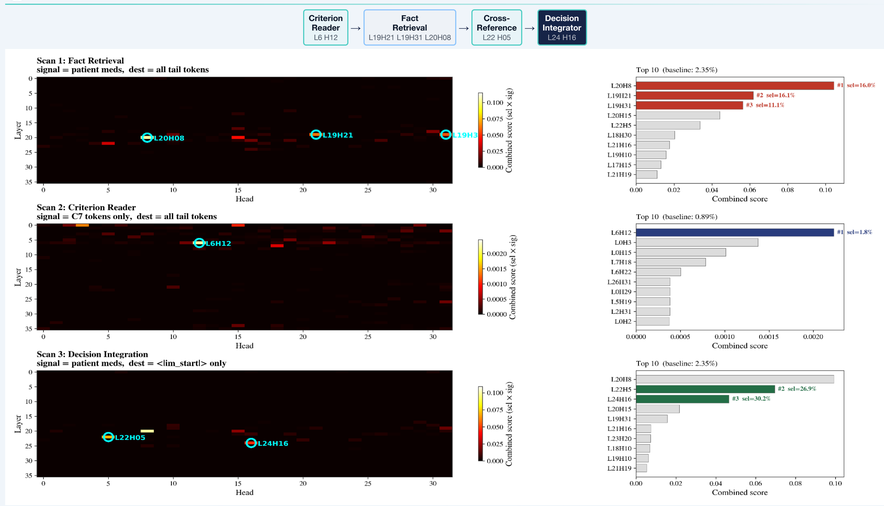
Probe vs. Frontier LLMs

Model	Iron	Osteo.	Kidney	Coag.	SLE	Mean
Linear Probe	95.5	87.0	86.0	93.0	88.0	**89.9%**
GPT-4o	96.0	92.0	100.0	100.0	96.0	96.8%
Claude Haiku	68.0	80.0	92.0	100.0	80.0	84.0%
GPT-4o-mini	84.0	44.0	68.0	72.0	84.0	70.4%

Probe outperforms Haiku and 4o-mini; approaches GPT-4o at far lower cost. Table by the author.

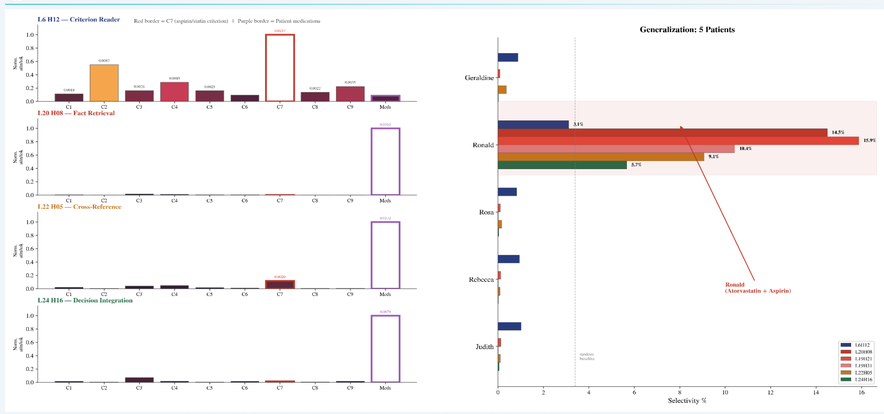
A linear probe on an open-source 8B model **outperforms Claude Haiku and GPT-4o-mini**, and approaches GPT-4o — at a fraction of the cost.

7 RESULTS: SIX-HEAD DECISION CIRCUIT



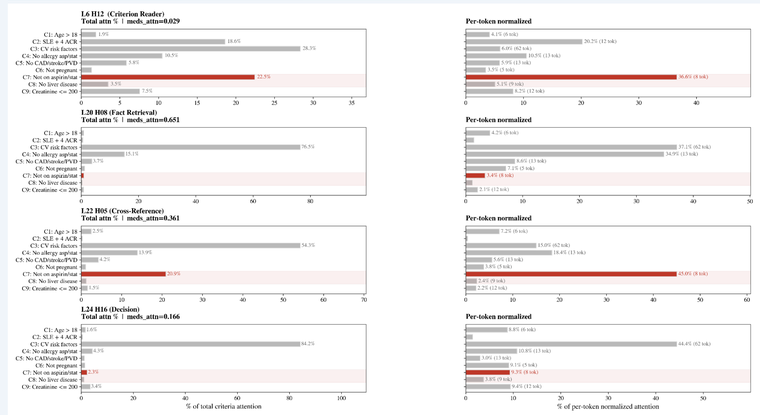
Three selectivity scans across 1,152 heads. Bright = high selectivity. Cyan = circuit heads. Figure created by the student researcher, 2026.

8 RESULTS: GENERALIZATION



*Circuit fires **only** for Ronald (Atorvastatin + Aspirin). All others at baseline. Confirms meaningful clinical signal. Figure created by the student researcher, 2026.*

9 RESULTS: HEAD SPECIALIZATION



Per-criterion attention for each circuit head. Red bars = C7 (aspirin/statin exclusion criterion). Figure created by the student researcher, 2026.

Key observation: The Criterion Reader (L6H12) broadly reads all criteria. The Fact Retriever (L20H08) focuses on patient medications. The Decision head (L24H16) concentrates **84% of attention** on the decisive criterion C3.

KEY RESULT

89.9% Accuracy

Linear Probe on Open-Source 8B Model

Outperforms Claude Haiku (84.0%) and GPT-4o-mini (70.4%)
while providing full mechanistic explainability

10 DISCUSSION

1. **Eligibility decisions are linearly encoded**
   Qwen3-8B encodes sufficient information about assessment correctness in its residual stream that a linear classifier can extract it without fine-tuning. Optimal probes consistently emerge at layers 21–24, indicating that decision-relevant features are consolidated in late layers.
2. **Per-trial specialization is necessary**
   Probe accuracy degrades monotonically with trial diversity (89.9% → 50.0%), confirming that each trial's eligibility logic occupies a distinct representational subspace. Per-trial probes are therefore the recommended deployment strategy.
3. **Mechanistic circuits enable native interpretability**
   The six-head decision circuit shows *how* the transformer processes eligibility decisions — which heads read criteria, retrieve patient facts, and integrate the final answer. This turns a black box into a glass box.

! LIMITATIONS

1. **Single architecture**
   Probing was limited to Qwen3-8B; generalizability to other model families remains untested.
2. **Low cross-trial transfer**
   Cross-trial probe transfer achieved only 55% accuracy, necessitating per-trial training.
3. **Correlational validation**
   Circuit validation is correlational; activation patching would provide stronger causal evidence.
4. **Small LLM baselines**
   Frontier LLM baselines were evaluated on 25 samples per trial; larger-scale comparisons are needed.

11 CONCLUSION

1. **Reliable verification**
   Linear probes on Qwen3-8B verify clinical trial eligibility with **89.9% mean accuracy** on held-out patients.
2. **Competitive performance**
   This 8B-parameter probe **surpasses Claude Haiku** (84.0%) **and GPT-4o-mini** (70.4%), approaching GPT-4o (96.8%) at substantially lower computational cost.
3. **Decision circuit identified**
   A sparse **six-head decision circuit** was identified with functionally distinct roles: criterion reading, fact retrieval, cross-referencing, and decision integration.
4. **Native explainability**
   Together, probes and circuits provide **transformer-native explainability** that may help build clinician trust in automated trial matching systems.

12 FUTURE WORK

1. **Multi-architecture validation**
   Extend probing and circuit discovery to Llama, Mistral, and other open-source families.
2. **Scaled evaluation**
   Increase LLM baseline sample sizes beyond 25 per trial to strengthen statistical comparisons.
3. **Causal verification**
   Apply activation patching to establish that the discovered circuit causally drives eligibility decisions.

REFERENCES

[1] Alain G, Bengio Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*. 2017.
[2] Burns C, Ye H, Klein D, Steinhardt J. Discovering latent knowledge in language models without supervision. *ICLR*. 2023.
[3] Conmy A, Mavor-Parker AN, Lynch A, Heimersheim S, Garriga-Alonso A. Towards automated circuit discovery for mechanistic interpretability. *NeurIPS*. 2023.
[4] Elhage N, Nanda N, Olsson C, et al. A mathematical framework for transformer circuits. *Anthropic Research*. 2021.
[5] Roberts K, Demner-Fushman D, Voorhees EM, et al. Overview of the TREC 2021 Clinical Trials Track. *TREC*. 2021.
[6] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172-180.
[7] Jain S, Wallace BC. Attention is not explanation. *Proc NAACL-HLT*. 2019:3543-3556.
[8] Weng C, Shah NH, Hripcsak G. Deep phenotyping: embracing complexity and temporality—towards scalability, portability, and interoperability. *J Biomed Inform*. 2020;105:103433.
[9] U.S. National Library of Medicine. ClinicalTrials.gov. Accessed 2025.

All graphics unless otherwise cited were created by the student researcher, 2026.