Supplementary Material 1

1. Ocular Motor Task Design:

All eye-movement assessments were performed using a custom iOS application recording via the iPhone TrueDepth infrared camera (60 Hz). Five canonical paradigms were administered:

1.1 Visually guided saccades

Two black "X" targets positioned 12.5 cm apart (≈20° visual angle at a viewing distance of ~35 cm) alternated every 2 seconds.

- **Vertical saccades:** participants followed the up–down alternation.
- **Horizontal saccades:** the phone was rotated 90° counterclockwise; the same task was repeated along the horizontal axis.

The application incorporated a leveling indicator to ensure consistent angular alignment.

1.2 Self-paced saccades (SPS)

Participants voluntarily alternated their gaze between the two fixed targets for 20 seconds.
Output: total number of saccades generated

1.3 Smooth pursuit (slow eye-tracking)

A black dot moved sinusoidally at 7°/s across the central visual field.

- Participants were instructed to "follow the black dot as smoothly as possible with your eyes."
- The predictable sinusoidal pattern enabled template-based trajectory comparison**.**

1.4 Anti-saccades

A peripheral stimulus appeared abruptly at predefined positions.

- Participants were instructed to look in the **opposite direction**.
- **Anti-saccade errors** = percentage of initial saccades made **towards** the stimulus.

Trajectory plots were used to confirm error labeling.

1.5 Fixation with distractors

Participants fixated a central “X” while peripheral images appeared briefly and unpredictably.

- Deviations > 2° from central fixation were scored as fixation losses, reflecting impaired inhibitory control.

1.6 Blink and head-thrust detection

Blink and head-thrust events were automatically flagged:

- **Blink:** 0 = absent, 1 = present (within window preceding saccade)
- **Head-thrust:** 0 = absent, 1 = present

These variables were used to mark data potentially impacted by artifacts.

2. Signal Processing Pipeline

Raw gaze-position traces were exported in *CSV* format and processed offline using a standardized pipeline:

1. **Blink removal and interpolation**
2. **Median filtering (3–9 samples)** to remove high-frequency noise
3. **Baseline correction** (detrending with 0.1 Hz high-pass)
4. **Velocity computation** via numerical derivative

Trajectories were kept in degrees of visual angle to preserve physiological interpretability.

3. Feature Extraction

The metrics provided to the LLMs belonged to **two distinct categories**:

3.1 Trajectory-based metrics (template matching)

Trajectory-based features were extracted for:

- **Vertical saccades**
- **Horizontal saccades**
- **Vertical smooth pursuit**
- **Horizontal smooth pursuit**

These tasks follow well-defined physiological patterns and therefore allow quantitative similarity assessment.

3.1.1 Physiological templates

- **Saccades:** modeled with a pulse–step function approximating canonical saccadic kinematics (fast ballistic rise + stable fixation).
- **Pursuit:** modeled using a sinusoidal trajectory

  $T(t)=A\cdot\sin f_0(2\pi ft+\phi)T(t)=A\cdot\sin(2\pi ft+\phi)$

Templates were scaled to each task's amplitude and sampling frequency.

3.1.2 Best-fit similarity algorithm

Each recorded trace **S(t)** was:

- amplitude-normalized,
- partially time-warped,
- aligned to the template **T(t)**.

Interpretation:

- **High AUC → near-normal physiology**
- **Low AUC → greater deviation**

Outputs:

- Vertical Saccades
- Horizontal Saccades
- Average Saccades
- Vertical Pursuit
- Horizontal Pursuit
- Average Pursuit
-

3.2 Event-based metrics

Tasks involving discrete events were quantified directly:

Metric	Meaning
AS Error Rate (%)	% incorrect initial saccades toward target
SPS count	Number of voluntary saccades in 20 s
Fixation deviations (%)	% epochs with drift > 2°

These features capture inhibitory control, executive function, and gaze stability, deficits well described in Huntington's disease.

Supplementary Material 2

1. Construction of LLM Input Vector

For each participant, the following values were concatenated in a structured numerical summary:

- **Vertical Saccades AUC**
- **Horizontal Saccades AUC**
- **Average Saccades AUC**
- **Vertical Pursuit AUC**
- **Horizontal Pursuit AUC**
- **Average Pursuit AUC**
- **AS Error Rate (%)**
- **SPS Count**
- **Fixation Deviations (%)**

No clinical data (UHDRS, TFC, MoCA) or diagnostic hints were included.

2. LLM Prompt Structure and Zero-Shot Inference

Each participant's metrics were fed **independently** to four LLMs:

- GPT-5.1
- DeepSeek R1
- Gemini
- Claude

2.1 Prompt Design

The prompt required the model to:

1. Interpret the physiological values.
2. Estimate a **probability of Huntington's disease (0–100%).**
3. Assign one categorical label:
    - "Likely HD"
    - "Uncertain"
    - "Likely Control"
4. Output results in JSON-like format.
5. Base inference **solely** on the numerical ocular motor metrics.
6. The exact prompt used for all models is included in the manuscript above:

PROMPT

"You are an AI research assistant helping in a study about Huntington's Disease (HD). Your task is to estimate the probability that a given case belongs to the HD group vs the control group, based ONLY on eye movement measurements. Higher values indicate greater oculomotor impairment and are more frequently found in HD.

Metrics provided:
- Vertical Saccades
- Horizontal Saccades
- Average Saccades
- Vertical Pursuit
- Horizontal Pursuit
- Average Pursuit
- Antisaccade Error Rate
- Self-paced Saccades
- Fixation Deviations (>2 degrees)
- Blink or head-thrust before saccade (yes/no)

YOUR TASK:

Estimate:

Probability of HD (0–100%)

Classification: "Likely HD", "Uncertain", or "Likely Control"

Base your reasoning only on these values.

Output strictly in this JSON-like format: { "HD_probability_percent": <number>, "classification": <string>, "reasoning": <short text>}"

2.2 Preventing diagnostic leakage

To ensure that disease detection was based purely on physiology:

- No labels/examples were provided.
- No clinical descriptions or keywords (“chorea”, “Huntington”, etc.) appeared in the input.
- Each participant was processed in isolation.
- Prompts were restricted to numerical values.

This guarantees **zero-shot** inference.

Supplementary Table 1

Ocular motor domain	Clinical score 0	Clinical score 1	Clinical score 2	Clinical score 3	Clinical score 4
Vertical saccades	0.86 [0.81–0.88]	1.31 [1.16–1.38]	1.57 [1.51–1.64]	2.22 [2.07–2.48]	4.13 [4.09–4.23]
Horizontal saccades	1.30 [1.01–1.82]	1.54 [1.26–1.66]	1.77 [1.70–1.88]	2.76 [2.39–3.08]	4.14 [4.12–4.94]
Vertical pursuit	0.67 [0.64–0.77]	1.10 [1.00–1.19]	1.28 [1.24–1.30]	1.87 [1.61–1.93]	2.68 [2.31–3.24]
Horizontal pursuit	0.64 [0.20–0.92]	1.26 [1.15–1.32]	1.50 [1.49–1.62]	1.79 [0]	3.19 [2.92–3.35]

Supplementary Table 1. Median [IQR] smartphone-derived ocular motor metrics across neurologist-rated clinical severity scores

Legend: Values are reported as median [interquartile range]. Higher values indicate worse ocular motor performance.

Supplementary Table 2

Patient	Diagnosis	GPT_class	GPT_prob %	DeepSeek_class	DeepSeek_prob %	Gemini_class	Gemini_prob %	Claude_class	Claude_prob %
1	HD	Likely HD	96	Likely HD	95	Likely HD	90	Likely HD	85
2	HC	Likely Control	4	Likely Control	5	Likely Control	2	Likely Control	5
3	HD	Likely HD	82	Likely HD	85	Likely HD	75	Likely HD	70
4	HD	Likely HD	88	Likely HD	80	Likely HD	80	Likely HD	65
5	HD	Uncertain	62	Uncertain	65	Uncertain	55	Uncertain	45
6	HC	Likely Control	18	Likely Control	10	Likely Control	1	Likely Control	15
7	HC	Likely Control	10	Likely Control	5	Likely Control	1	Likely Control	10
8	HC	Likely Control	20	Likely Control	20	Likely Control	3	Likely Control	12
9	HD	Likely HD	85	Likely HD	90	Likely HD	85	Likely HD	75
10	HD	Uncertain	48	Uncertain	60	Uncertain	60	Uncertain	40
11	HD	Likely Control	22	Uncertain	30	Uncertain	30	Likely Control	8
12	HD	Likely Control	28	Uncertain	40	Uncertain	45	Likely Control	25
13	HC	Likely Control	12	Likely Control	5	Likely Control	1	Likely Control	5
14	HD	Likely Control	24	Uncertain	25	Uncertain	20	Likely Control	8
15	HD	Likely HD	92	Likely HD	98	Likely HD	95	Likely HD	90
16	HC	Likely Control	18	Uncertain	25	Uncertain	15	Likely Control	10
17	HD	Likely HD	97	Likely HD	99	Likely HD	99	Likely HD	95
18	HD	Likely Control	18	Likely Control	5	Likely Control	2	Likely Control	15
19	HD	Uncertain	42	Uncertain	70	Uncertain	65	Uncertain	35
20	HC	Likely Control	12	Likely Control	2	Likely Control	1	Likely Control	10
21	HC	Likely Control	8	Likely Control	1	Likely Control	1	Likely Control	5

22	HC	Uncertain	28	Uncertain	60	Uncertain	10	Likely Control	8
23	HC	Likely Control	20	Uncertain	55	Uncertain	5	Likely Control	5
24	HC	Likely Control	15	Uncertain	25	Uncertain	35	Likely Control	12
25	HC	Likely Control	22	Likely Control	15	Likely Control	3	Likely Control	18
26	HC	Likely Control	26	Uncertain	60	Uncertain	15	Likely Control	20

Supplementary Table 2: AI-derived probability classifications for HD vs. control across four large language models. Legend This table reports the categorical classifications ("Likely HD", "Uncertain", "Likely Control") and continuous probability estimates (0–100%) generated by four independent large language models (GPT-5.1, DeepSeek R1, Gemini, and Claude) for all participants in the study (HD = Huntington's disease; HC = healthy controls). Models were prompted in zero-shot mode using only smartphone-derived oculomotor metrics without access to clinical or diagnostic information. Higher probability values indicate greater oculomotor impairment consistent with HD physiology. Class labels reflect model-specific decision rules defined in the structured prompt. **Abbreviations:** HD = Huntington's disease; HC = healthy control.