Accuracy Of a GPT-Based Gait Report Interpreter: Matching Expert Panel Performance

Hitav Someshwar[1], Mitesh Chawda[2], Abhishek Jaroli[3]
[1]*Assistant Professor, Early Intervention and Rehabilitation Centre for Children, TNMC & BYL Nair Ch. Hospital, Mumbai.*
[2]*Consultant Pediatrician, EKTA INSTITUTE OF CHILD HEALTH*
[3]*Jr consultant orthopaedic dept, KBBH govt hospital Bandra West Mumbai*

***Abstract*—Background Instrumented gait analysis generates complex datasets requiring expert interpretation. This process is time-consuming and subject to inter-rater variability. Large language models (LLMs) offer potential as decision-support tools, but their accuracy must be validated. We evaluated a GPT-based gait report interpreter against an expert clinical panel.**

Methods: We retrospectively analysed 150 de-identified gait reports (2019–2024) spanning cerebral palsy, stroke, Guillain-Barré syndrome, brachial plexus injury, and orthopaedic cases. An expert panel of four clinicians—two physiotherapists (5- and 10-years' experience), one orthopaedic surgeon (25 years), and one rehabilitation physician (20 years)—independently annotated deviations, severity, and biomechanical drivers. Consensus served as the reference. The GPT-based interpreter (frozen v1.0 prompt) generated blinded outputs. Accuracy was assessed with sensitivity, specificity, multi-label F1 score, and weighted Cohen's κ.

Results: The panel identified 1,120 deviations (mean 7.5/case). The model achieved 87% sensitivity, 90% specificity, and a multi-label F1 of 0.84, non-inferior to the median expert (0.82; NI margin −0.05). For severity grading, weighted κ was 0.62 (95% CI 0.58–0.66), within the range of individual clinicians (0.59–0.68). Agreement on biomechanical drivers reached κ = 0.58 (95% CI 0.52–0.64). Errors were mostly minor (side mislabelling, gait phase misinterpretation, mild underestimation of crouch). No critical errors were recorded.

Conclusion: The GPT-based interpreter identified deviations and their causes with accuracy equivalent to experienced clinicians. Its performance was within the expert range, supporting its potential as a safe, supervised decision-support tool in gait laboratories. Prospective trials are needed to assess workflow efficiency and decision impact.

***Index Terms*—gait analysis; artificial intelligence; GPT; rehabilitation; biomechanics**

I. INTRODUCTION

Instrumented gait analysis has become a cornerstone in the assessment of children and adults with complex movement disorders. Motion capture systems, force plates, and surface electromyography generate hundreds of variables for each walking trial [1,2]. These include spatiotemporal parameters such as cadence and step length, kinematic curves describing joint motion, kinetic data capturing joint moments and powers, and EMG traces reflecting muscle activation patterns. The richness of these datasets allows for precise characterization of gait pathology, but the process of interpretation is far from straightforward [3].

Traditionally, experienced gait analysts and clinicians synthesize these streams of information into a structured report. This requires not only technical familiarity with the output formats but also the ability to integrate biomechanical reasoning with clinical context. For example, identifying a crouch gait pattern involves recognizing excessive knee flexion instance, linking it to altered hip and ankle mechanics, and then considering whether the driver is spasticity, weakness, or skeletal deformity [1,2,4]. The expertise needed to perform this integration typically develops over years of practice.

The challenge is twofold. First, interpretation is time-consuming: constructing a report can take an expert thirty minutes or more, and even longer for junior staff. Second, interpretation is variable. Different

clinicians may highlight different deviations or attribute them to different underlying mechanisms [2,5]. This variability can affect downstream decision-making, including whether to recommend physiotherapy, orthotic intervention, botulinum toxin injection, or surgery.

Artificial intelligence, and more recently large language models (LLMs), offer a potential solution [6,7]. These models are designed to process complex text and generate structured, human-like outputs. When applied to gait analysis, an LLM can take the numerical and graphical outputs of motion capture and produce a narrative that resembles what a clinician might write [4]. Beyond simple restatement of numbers, the promise lies in identifying clinically meaningful deviations and linking them to plausible biomechanical causes.

However, the adoption of AI in a clinical setting requires rigorous validation. It is not enough for a model to produce convincing language; its statements must be factually accurate and clinically sound [8]. This means showing that the model can detect deviations as reliably as human experts and that its explanations do not introduce hallucinations or critical errors. To date, evidence in this space has been limited, with most studies focusing on small pilot sets or narrow patient groups [4,7].

The present study addresses this gap by evaluating the accuracy of a GPT-based gait report interpreter. Specifically, we asked whether the model could identify key gait deviations and their likely causes as accurately as an expert clinical panel. By benchmarking performance against consensus labels across a diverse patient cohort, we aimed to determine whether the model meets the standard required for use as a decision-support tool in gait laboratories.

II. METHODS

We conducted a retrospective evaluation of 150 de-identified gait reports collected in a tertiary gait laboratory. The cohort included children and adults across a spectrum of diagnoses: cerebral palsy, stroke, Guillain-Barré syndrome, traumatic brain injury, and orthopaedic pathologies. Each report contained standard spatiotemporal parameters, three-dimensional kinematics, kinetics, and surface EMG summaries.

AN EXPERT PANEL OF FOUR CLINICIANS INDEPENDENTLY ANNOTATED EVERY CASE:

- Physiotherapist A with 5 years' experience in clinical gait analysis.
- Physiotherapist B with 10 years' experience in gait analysis and neurorehabilitation.
- Orthopaedic surgeon with 25 years' experience in surgical and conservative management of gait disorders.
- Rehabilitation physician with 20 years' experience in neurorehabilitation and gait interpretation.

EACH CLINICIAN DOCUMENTED:

1. Presence or absence of predefined gait deviations, including equinus, crouch gait, stiff-knee gait, recurvatum knee, foot drop, genu valgum, genu varum, excessive internal or external foot progression, and Trendelenburg gait.
2. Severity grading for each deviation (none, mild, moderate, severe).
3. Likely biomechanical driver, categorized as kinematic (joint motion abnormality), kinetic (abnormal forces or powers), or neuromuscular (e.g., spasticity, weakness, abnormal activation).

Consensus labels were derived through blinded adjudication: when at least three of the four clinicians agreed, that judgment became the reference standard. Cases without majority agreement were discussed in a second round of blinded review until consensus was reached.

The GPT-based gait report interpreter (frozen v1.0 prompt and temperature settings) was then applied to the same 150 reports. Model outputs were anonymized, version-controlled, and scored blindly against consensus.

PERFORMANCE METRICS INCLUDED:

- Per-feature sensitivity and specificity for deviation detection.
- Multi-label F1 score for deviation identification at the case level.
- Weighted Cohen's kappa (κ) for agreement on severity grading.
- κ agreement on biomechanical driver classification.
- Error taxonomy, including left/right mislabelling, phase misinterpretation, and under/over-estimation of severity.

III. RESULTS

Across the 150 gait reports, the expert panel identified a total of 1,120 distinct gait deviations (mean 7.5 per case). The GPT-based interpreter was able to detect deviations with a sensitivity of 87% and a specificity of 90% compared to consensus. Its multi-label F1 score was 0.84, which was statistically non-inferior to the median expert (0.82) under the pre-specified margin of −0.05.

SEVERITY GRADING

For deviation severity (none, mild, moderate, severe), the model achieved a weighted $\kappa = 0.62$ (95% CI 0.58–0.66), representing substantial agreement with consensus. This was comparable to the κ values between individual clinicians and the panel:

- Physiotherapist A (5 years' experience): $\kappa = 0.59$
- Physiotherapist B (10 years' experience): $\kappa = 0.64$
- Orthopaedic surgeon (25 years' experience): $\kappa = 0.68$
- Rehabilitation physician (20 years' experience): $\kappa = 0.61$

Thus, the model's agreement was within the expert range, outperforming the less experienced physiotherapist and closely matching the rehabilitation physician.

BIOMECHANICAL DRIVER CLASSIFICATION

When attributing deviations to likely biomechanical drivers (kinematic, kinetic, neuromuscular), the model reached $\kappa = 0.58$ (95% CI 0.52–0.64) against consensus. This was similar to inter-rater reliability across clinicians:

- Physiotherapist A: $\kappa = 0.55$
- Physiotherapist B: $\kappa = 0.60$
- Orthopaedic surgeon: $\kappa = 0.63$
- Rehabilitation physician: $\kappa = 0.57$

This shows the model not only aligned with consensus but also tracked closely with the reasoning patterns of senior clinicians, particularly the orthopaedic surgeon and senior physiotherapist.

ERROR ANALYSIS

A structured audit of model misclassifications revealed:

- 41 cases of severity misestimation (most commonly under-calling moderate crouch as mild).
- 27 cases of left/right limb mislabelling.
- 19 instances of phase timing misinterpretation (e.g., terminal stance vs pre-swing).
- 12 cases of ambiguous driver attribution (kinematic vs neuromuscular).

Importantly, no errors were classified as "critical", defined as statements that could plausibly lead to unsafe clinical recommendations if left uncorrected.

COMPARATIVE PERSPECTIVE

- The model's accuracy was higher than the less experienced physiotherapist, particularly for crouch gait and foot progression deviations.
- It matched the rehabilitation physician in deviation detection and exceeded them slightly in severity agreement.
- It remained slightly below the orthopaedic surgeon for complex multi-planar deviations (e.g., combined valgus + rotational abnormalities), but within non-inferiority margins.

IV. DISCUSSION

This study evaluated the accuracy of a GPT-based gait report interpreter across 150 gait analyses, using an expert panel of four clinicians as the reference. The model achieved high sensitivity and specificity for detecting deviations, substantial agreement on severity grading, and acceptable agreement on biomechanical driver attribution. Importantly, its performance was within the range of human experts, outperforming a less experienced physiotherapist and aligning closely with senior clinicians.

CLINICAL IMPLICATIONS

The finding that the model's deviation detection was non-inferior to expert consensus is significant. Previous work has shown that gait analysis interpretation is not only time-consuming but also subject to inter-rater variability, especially when different clinical backgrounds are involved [1,2]. Our data suggest that a GPT-based tool could help standardize interpretations by providing consistent, structured outputs. This has particular relevance in multidisciplinary teams, where physiotherapists, surgeons, and rehabilitation physicians may emphasize different aspects of a gait pattern [3].

The model's agreement on severity and biomechanical drivers also mirrors existing evidence that even among experts, κ values for these judgments rarely exceed 0.65 [2,4]. The tool, therefore, sits squarely in the expert range, offering a level of reproducibility that could be valuable for both clinical decision-making and longitudinal tracking of patients. For junior staff, the interpreter may function as a training adjunct, exposing them to expert-level reasoning in a structured, repeatable way [5].

ERROR PROFILE

Our error analysis revealed predictable weaknesses: side mislabelling, misjudging gait phase, and minor underestimation of crouch severity. These are comparable to common human interpretation errors reported in prior work [1,4]. Crucially, no critical errors were identified, which supports the model's safety for supervised clinical use. However, the presence of minor misclassifications underscores the need for mandatory human oversight, particularly in high-stakes cases where surgical recommendations are considered [6,7].

BROADER CONTEXT OF AI IN MEDICINE

Large language models have been increasingly explored as clinical decision-support tools. Systematic reviews show they can achieve expert-level performance in text-based diagnostic and triage tasks, but issues of factual fidelity and hallucination remain barriers to deployment [7,8]. Our results demonstrate that when carefully constrained (frozen prompts, controlled input data, structured scoring), LLMs can achieve robust analytical and clinical validity in a complex biomechanical domain.

LIMITATIONS AND FUTURE WORK

This study has limitations. First, it was retrospective and based in a single tertiary gait lab; external validation in other centres, with different hardware and patient populations, is needed to establish generalizability [2,6]. Second, the consensus process, while rigorous, reflects the biases of the specific expert panel. Third, we did not assess workflow efficiency or decision impact—these are the focus of Phase C and D. Future work should include prospective randomized trials to measure whether the model reduces reporting time, enhances confidence, and improves clinical decisions. Safety audits and bias analyses across age, diagnosis, and assistive device subgroups will also be crucial before clinical integration.

V. CONCLUSION

Our findings show that a GPT-based gait interpreter can identify deviations and their likely causes as accurately as an expert panel, with performance comparable to senior clinicians. This establishes a strong foundation for its use as a decision-support tool, provided human oversight remains mandatory. With further validation, the model has potential to enhance efficiency, reduce variability, and improve access to expert-level gait interpretation.

REFERENCES

[1] King SL, Barton GJ, Ranganath LR. Interpreting sources of variation in clinical gait analysis. *Gait Posture*. 2017; 52:1-4.

[2] Leary E, et al. Revisiting sources of variability in gait analysis. *Gait Posture*. 2025; (in press).

[3] Schwartz MH, Georgiadis AG, et al. Evidence Based Gait Analysis Interpretation Tools (EB-GAIT). *PLoS One*. 2025; (expected).

[4] Hausdorff JM. Gait variability: methods, modelling and meaning. *J Neuro Engineering Rehabil*. 2005; 2:19.

[5] Özateş ME, Yaman A, et al. Identification and interpretation of gait analysis features and foot conditions by explainable AI. *Sci Rep*. 2024; 14:5998.

[6] Labkoff S, et al. Recommendations for AI-enabled clinical decision support. *J Am Med Inform Assoc*. 2024;31(11):2730-40.
[7] Shool S, et al. A systematic review of large language model (LLM) use in medicine. *BMC Med Inform Decis Mak*. 2025; 25:54.
[8] Hager P, et al. Evaluation and mitigation of the limitations of large language models for clinical decision-making. *Nat Med*. 2024; 30:1461-70.