

Can Artificial Intelligence Calculate Disability Rates? A Comparative Study of Three Large Language Models Against Expert Assessment

Yasin Etli¹, Erhan Kartal¹, Mahmut Aşirdizer²

¹Department of Forensic Medicine, Faculty of Medicine, Van Yüzüncü Yıl University, Van, Türkiye

²Department of Forensic Medicine, Faculty of Medicine, Bahçeşehir University, İstanbul, Türkiye

Abstract

Introduction: Disability impairment rating is a critical medicolegal function requiring precise integration of clinical findings with regulatory tables. Large language models (LLMs) have shown promising capabilities in medicine; however, their ability to perform quantitative disability assessments has not been evaluated. This study aimed to assess the agreement between three commercial LLMs and expert-determined disability ratings.

Materials and Methods: One hundred standardized fictional clinical cases were constructed representing the spectrum of medicolegal disability assessments. Each case was evaluated by a forensic medicine specialist using the Turkish Regulation on Disability Assessment for Adults (2019) as the reference standard. Three LLMs—GPT-5, Gemini 3 Pro, and Claude Opus 4.6—were tested using a zero-shot prompt. A three-model ensemble was also examined. Agreement was assessed using intraclass correlation coefficients (ICC), Bland-Altman analysis, and mean absolute error (MAE).

Results: All three LLMs demonstrated good agreement with the reference standard, with ICC values of 0.875 (Gemini 3 Pro), 0.848 (Claude Opus 4.6), and 0.814 (GPT-5). MAE ranged from 5.1 to 6.0 percentage points. No significant difference was found among models (Friedman $p=0.790$). The three-model ensemble achieved the highest ICC (0.890) and lowest MAE (4.7 points). Accuracy declined substantially for rates $\geq 21\%$ (MAE 9.1–11.5). Agreement for temporary work disability duration was moderate (ICC 0.701–0.750), while caregiver need showed poor-to-moderate agreement (ICC 0.497–0.585).

Conclusion: Commercial LLMs demonstrate good overall agreement with expert disability ratings but show declining accuracy for complex multi-system cases. A multi-model ensemble approach improves reliability. LLMs may serve as supportive tools in medicolegal disability assessment, although expert oversight remains essential.

Key words: Artificial intelligence; disability evaluation; forensic medicine; large language models; reproducibility of results; medicolegal assessment.

Introduction

Large language models (LLMs) are artificial intelligence systems capable of generating human-like textual responses through processing vast amounts of text data (1). Since ChatGPT's release in November 2022, these models have been rapidly integrated into healthcare and medical research (2). LLMs can achieve passing scores on medical licensing examinations, with a pooled accuracy of 81.8% for ChatGPT-4 (3), and a randomized clinical trial reported that an LLM alone outperformed physicians in diagnostic reasoning (4). The application of LLMs in specialized medical fields has yielded variable results. A systematic review of 519 studies concluded that existing evaluations are fragmented and insufficient in covering diverse specialties (5), with most investigations using fewer than 100 test

items (6). LLM accuracy has also been shown to vary across levels of clinical complexity (7). In forensic medicine, LLMs have been recognized as a new frontier with both opportunities and challenges (8). A Turkish study evaluating ChatGPT-4 across 366 medicolegal cases found the model could aid forensic medicine assistants, although discrepancies existed between AI-generated and expert-prepared reports (9). A recent benchmarking study of eleven multimodal LLMs on 847 forensic questions reported emerging potential for structured assessments but persistent limitations in complex scenarios (10). Disability assessment is a critical function of forensic medicine that directly influences compensation in personal injury litigation. In Turkey, disability ratings for adults are determined by the "Regulation on Disability Assessment for Adults" (Official Gazette, 20 February 2019, No.

*Corresponding Author: Yasin Etli Department of Forensic Medicine, Faculty of Medicine Van Yüzüncü Yıl University Dursun Odabaşı Medical Center 65080, Tuşba, Van, Türkiye Email: yasinetli@yyu.edu.tr Orcid: Yasin Etli [0000-0002-7369-6083](https://orcid.org/0000-0002-7369-6083), Erhan Kartal [0000-0003-2459-7756](https://orcid.org/0000-0003-2459-7756), Mahmut Aşirdizer [0000-0001-7596-5892](https://orcid.org/0000-0001-7596-5892)

Received: 05.03.2026, Accepted: 06.05.2026



Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

30692), which provides standardized impairment tables and the Balthazard formula for combining multiple impairments (11). This assessment requires integrating clinical findings, imaging data, and regulatory tables—demanding both medical expertise and precise numerical calculation. Despite growing interest in AI-assisted medicolegal documentation, no study has specifically evaluated the ability of LLMs to calculate disability ratings using national regulatory frameworks. This study evaluated three leading commercial LLM-based chatbots—Gemini 3 Pro (Google), ChatGPT Plus (GPT-5; OpenAI), and Claude AI (Opus 4.6; Anthropic)—in calculating disability impairment ratings, temporary work disability duration, and caregiver need duration using standardized cases based on the Turkish regulation. Comparing AI outputs against expert-determined values across 100 cases, this is the first systematic assessment of LLM accuracy in quantitative medicolegal disability evaluation.

Materials And Methods

Study design: This was a cross-sectional agreement study comparing disability impairment ratings generated by three commercial LLM-based chatbots—individually and as a three-model ensemble—against expert-determined reference values across 100 standardized cases. The study was reported in accordance with the GRRAS guidelines (12).

Ethics statement: All clinical cases were fictional, purpose-constructed narratives; therefore, ethical committee approval and informed consent were not required. The use of AI tools is declared in accordance with Van Medical Journal policy.

Case construction: One hundred fictional clinical cases were constructed by a forensic medicine specialist (Y.E.) to represent the clinical spectrum encountered at the Forensic Medicine Outpatient Clinic of Dursun Odabaş Medical Center in terms of age distribution, injury mechanisms, affected body systems, and disability rate distribution. Each case included: demographic information (age, sex), injury mechanism and date, medical history and surgical interventions, physical examination findings with range-of-motion measurements, imaging results, and current clinical status. The distribution of injury mechanisms was as follows: traffic accidents (n=89), other accidents (n=6), assault (n=2), occupational injuries (n=2), and falls (n=1). The affected body systems included the

musculoskeletal system (lower extremity, upper extremity, and spine), neurological system, thorax, internal organs, eyes, ear-nose-throat, and psychiatric conditions. Seventy cases involved multiple impairments requiring the Balthazard combined rating formula, while 30 cases involved a single impairment.

Reference standard (Gold Standard): Each case was independently evaluated by a forensic medicine specialist using the "Regulation on Disability Assessment for Adults" (Official Gazette, February 20, 2019, No. 30692) (11). Three outcome measures were determined for each case: (1) permanent disability impairment rate (%), (2) temporary work disability duration (days), and (3) caregiver need duration (days). The impairment tables, figures, and the Balthazard formula specified in Annex-2 of the regulation were applied. These expert-determined values served as the reference standard for all comparisons. The distribution of reference disability rates was as follows: 0% (n=9), 1–10% (n=41), 11–20% (n=24), and ≥21% (n=26), with a mean of $16.0 \pm 16.0\%$ (range: 0–100%).

Large language model evaluation: Three commercial LLM-based chatbots were evaluated using paid subscription tiers: ChatGPT Plus with GPT-5 (OpenAI, USA), Google AI Pro with Gemini 3 Pro (Google, USA), and Claude Pro with Claude Opus 4.6 (Anthropic, USA). All evaluations were conducted in February 2026 via the web interfaces of each platform. A new conversation session was initiated for each batch to prevent context carryover. A standardized zero-shot prompt was developed that included: (a) a system role definition assigning the model the persona of an experienced forensic medicine expert, (b) the complete text of the relevant regulation tables and classification criteria from Annex-2, (c) the Balthazard formula with a worked example, and (d) explicit instructions to provide a single integer value for each of the three outcome measures without disclaimers or ranges. The cases were presented in batches of 10, and the models were instructed to output a structured summary table at the end of each batch. The same prompt template and case text were used identically across all three models.

Outcome measures: The primary outcome was the disability impairment rate (%). Secondary outcomes were temporary work disability duration (days) and caregiver need duration (days). For one case with lifelong caregiver need and one case with indeterminate temporary work disability,

Table 1: Descriptive Statistics of Disability Impairment Rate, Temporary Work Disability Duration, and Caregiver Need Duration

Source	Disability Impairment Rate (%)					
	n	Mean $\pm$ SD	Median	IQR (Q1–Q3)	Min	Max
Expert (Reference)	100	16.0 $\pm$ 16.1	10.5	5–21	0	100
Gemini 3 Pro	100	14.9 $\pm$ 17.1	10.0	3–19	0	96
GPT-5	100	14.4 $\pm$ 16.3	8.0	5–20	0	100
Claude Opus 4.6	100	15.7 $\pm$ 17.0	11.5	3–22	0	100
Three-Model Ensemble	100	15.0 $\pm$ 16.0	11.0	5–18	0	99
Source	Temporary Work Disability Duration (days)					
	n	Mean $\pm$ SD	Median	IQR (Q1–Q3)	Min	Max
Expert (Reference)	99	155.1 $\pm$ 111.1	180	90–225	15	540
Gemini 3 Pro	99	165.7 $\pm$ 95.7	180	90–270	15	365
GPT-5	99	152.1 $\pm$ 93.3	180	90–225	10	365
Claude Opus 4.6	99	174.9 $\pm$ 91.9	180	105–270	15	365
Three-Model Ensemble	99	164.3 $\pm$ 87.1	180	98–230	13	363
Source	Caregiver Need Duration (days)					
	n	Mean $\pm$ SD	Median	IQR (Q1–Q3)	Min	Max
Expert (Reference)	95	64.6 $\pm$ 53.9	60	26–90	0	180
Gemini 3 Pro	99	47.2 $\pm$ 46.4	45	15–60	0	365
GPT-5	100	47.6 $\pm$ 57.9	30	0–60	0	365
Claude Opus 4.6	99	54.1 $\pm$ 46.8	45	30–90	0	360
Three-Model Ensemble	95	45.3 $\pm$ 32.4	50	18–70	0	150

IQR, interquartile range; SD, standard deviation.

these values were treated as missing in the corresponding analyses.

Ensemble model approach: In addition to evaluating each LLM individually, an ensemble approach was examined in which the arithmetic mean of the three models' outputs was calculated for each case and each outcome measure. The ensemble value for disability impairment rate was rounded to the nearest integer for categorical accuracy comparisons. This approach was based on the rationale that aggregating predictions from multiple independent models may reduce individual model errors and yield more robust estimates—a principle well established in machine learning as ensemble averaging (13).

Statistical analysis: Descriptive statistics were presented as mean $\pm$ standard deviation and median (interquartile range) for continuous variables, and as frequencies and percentages for categorical variables. The Shapiro-Wilk test was used to assess the normality of difference distributions. Agreement between each LLM and the reference standard was evaluated using the intraclass correlation coefficient (ICC), with a two-way mixed-effects model and absolute agreement for single measures [ICC(3,1)] (14). ICC values were interpreted as: <0.50 poor, 0.50–0.75 moderate, 0.75–0.90 good, and >0.90 excellent (15). Spearman rank correlation

coefficients (ρ) were calculated for monotonic association, as the disability rate distributions were non-normally distributed. Systematic bias was quantified using the Bland-Altman method: the mean difference (bias), standard deviation of differences, and 95% limits of agreement (LoA = bias $\pm$ 1.96 $\times$ SD) were calculated and plotted for each AI model (16). Error magnitude was assessed using mean absolute error (MAE) and root mean square error (RMSE). Categorical accuracy was reported as the proportion of cases where the AI-generated value exactly matched the reference value (exact agreement) and the proportion within ± 5 and ± 10 percentage points of the reference. The Friedman test was applied to absolute errors to compare the three AI models, with post-hoc Wilcoxon signed-rank tests and Bonferroni correction when significant. The Wilcoxon signed-rank test was also used to compare the absolute errors of the ensemble against each individual model. Subgroup analyses were conducted by stratifying cases according to reference disability rate categories (0%, 1–10%, 11–20%, $\geq 21\%$) to evaluate whether AI performance varied across the severity spectrum. The same agreement metrics (ICC, MAE, Bland-Altman) were calculated for the ensemble predictions. An additional subgroup analysis was performed by categorizing cases as single impairment (n=30) or

Table 2: Agreement and error metrics for disability impairment rate between AI models and expert reference.

Model	ICC (3,1) [95% CI]	Spearman r _s	MAE (point)	RMSE (point)	Bias ± SD (points)	95% LoA	Exact Agreement n (%)	±5 Points n (%)	±10 Points n (%)
Gemini 3 Pro	0.875 [0.821–0.915]	0.833	5.1	8.3	-1.1 ± 8.3	-17.4 to 15.1	24 (24%)	73 (73%)	84 (84%)
GPT-5	0.814 [0.739–0.873]	0.788	6.0	9.9	-1.6 ± 9.8	-20.8 to 17.6	31 (31%)	63 (63%)	78 (78%)
Claude Opus 4.6	0.848 [0.781–0.894]	0.821	5.9	9.1	-0.4 ± 9.2	-18.3 to 17.6	21 (21%)	65 (65%)	81 (81%)
Three-Model Ensemble	0.890 [0.843–0.926]	0.870	4.7	7.5	-1.1 ± 7.5	-15.7 to 13.6	20 (20%)	75 (75%)	85 (85%)

ICC, intraclass correlation coefficient; CI, confidence interval; MAE, mean absolute error; RMSE, root mean square error; LoA, limits of agreement.

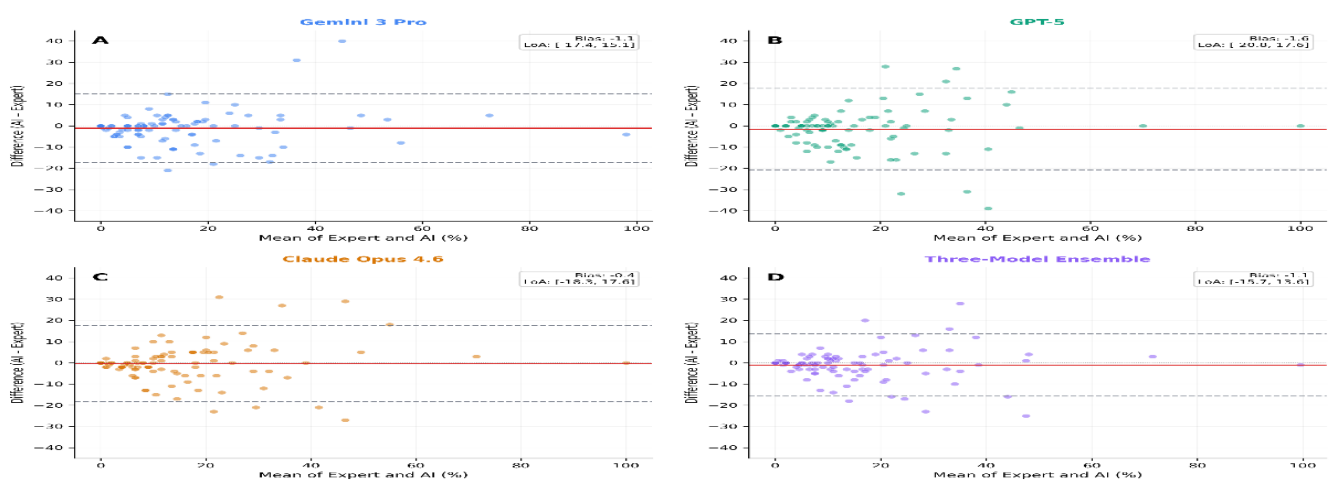


Figure 1: Bland-Altman plots showing agreement between AI-generated and expert-determined disability impairment rates. Panels A–C represent individual models (Gemini 3 Pro, GPT-5, and Claude Opus 4.6, respectively), and Panel D represents the three-model ensemble. The solid red line indicates mean difference (bias), and dashed gray lines indicate 95% limits of agreement (bias ± 1.96 SD)

Table 3: Subgroup Analysis of Mean Absolute Error and Bias by Disability Severity Category

Category	n	Gemini 3 Pro MAE	Gemini 3 Pro Bias	GPT-5 MAE	GPT-5 Bias	Claude Opus 4.6 MAE	Claude Opus 4.6 Bias	Ensemble MAE	Ensemble Bias
0%	9	0.0	+0.0	0.0	+0.0	0.2	+0.2	0.1	+0.1
1–10%	41	3.1	-0.4	2.8	+0.8	3.5	+0.9	2.5	+0.4
11–20%	24	5.3	-1.9	7.8	-3.8	6.5	+0.2	5.2	-1.9
≥21%	26	9.9	-2.0	11.5	-3.9	11.0	-3.1	9.1	-3.0
Overall	100	5.1	-1.1	6.0	-1.6	5.9	-0.4	4.7	-1.1

MAE, mean absolute error. Bias = AI prediction minus expert reference (negative values indicate underestimation).

multiple impairment requiring the Balthazard combined rating formula (n=70) to evaluate whether the computational complexity of cumulative calculations affected AI performance. The Mann-Whitney U test was used to compare absolute errors between these two groups. All statistical analyses were performed using Python 3.12 with SciPy 1.17, NumPy 2.4, and Pandas 3.0 libraries. Figures were generated using Matplotlib 3.10. A two-tailed p-value of <0.05 was considered statistically significant.

Results

All 100 cases were successfully evaluated by each LLM, with complete data for the primary outcome. For temporary work disability duration and caregiver need duration, 99 and 95 cases were evaluable, respectively. Descriptive statistics are summarized in Table 1.

Disability impairment rate: The expert-determined disability rates ranged from 0% to 100% (mean $16.0 \pm 16.1\%$, median 10.5%). The agreement and error metrics for each model and the three-model ensemble are presented in Table 2. All three models showed good agreement with the reference standard, with ICC values from 0.814 (GPT-5; 95% CI: 0.739–0.873) to 0.875 (Gemini 3 Pro; 95% CI: 0.821–0.915), and Spearman coefficients of 0.788–0.833 (all $p < 0.001$). Gemini 3 Pro had the lowest MAE (5.1 points); GPT-5 had the highest exact agreement (31%). All models showed slight negative bias, indicating a tendency to underestimate disability rates (Table 2). Bland-Altman limits of agreement (Figure 1) were narrowest for Gemini 3 Pro (−17.4 to +15.1) and widest for GPT-5 (−20.8 to +17.6).

Ensemble model performance: The ensemble yielded the highest ICC (0.890; 95% CI: 0.843–0.926), approaching excellent agreement. It also achieved the lowest MAE (4.7), lowest RMSE (7.5), narrowest limits of agreement (−15.7 to +13.6), and highest proportion within ± 5 pp (75%) (Table 2). The ensemble significantly reduced absolute errors compared with Claude Opus 4.6 ($p < 0.001$) and GPT-5 ($p = 0.028$); improvement over Gemini 3 Pro approached significance ($p = 0.065$).

Inter-Model comparison: The Friedman test revealed no statistically significant difference in absolute error magnitudes among the three individual models ($\chi^2 = 0.47$, $p = 0.790$), indicating comparable overall performance.

Subgroup analysis: Performance varied substantially across disability categories (Table 3). For 0% cases, all models achieved near-perfect

accuracy ($MAE \leq 0.2$). MAE in the 1–10% range was 2.5 (ensemble) to 3.5 (Claude Opus 4.6). Accuracy declined progressively with severity: in the $\geq 21\%$ category, MAE reached 9.1 (ensemble) to 11.5 (GPT-5), with all models showing negative bias and systematic underestimation (Figure 2). These error patterns are summarized in Figure 3, which presents the distribution of absolute errors across the individual models and the ensemble (A), the proportion of predictions falling within increasing tolerance thresholds (B), and mean absolute error stratified by disability rate category (C).

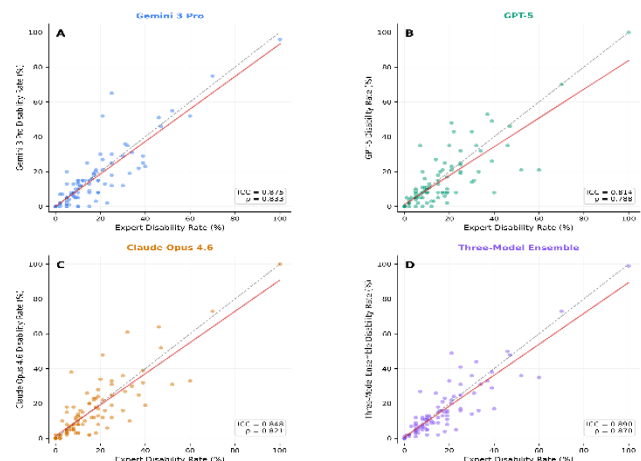


Figure 2: Correlation scatter plots between expert-determined and AI-generated disability impairment rates. Panels A–C represent individual models (Gemini 3 Pro, GPT-5, and Claude Opus 4.6, respectively), and Panel D represents the three-model ensemble. The dashed black line indicates perfect agreement (identity line), and the solid red line indicates the linear regression fit. ICC and Spearman ρ values are shown in each panel.

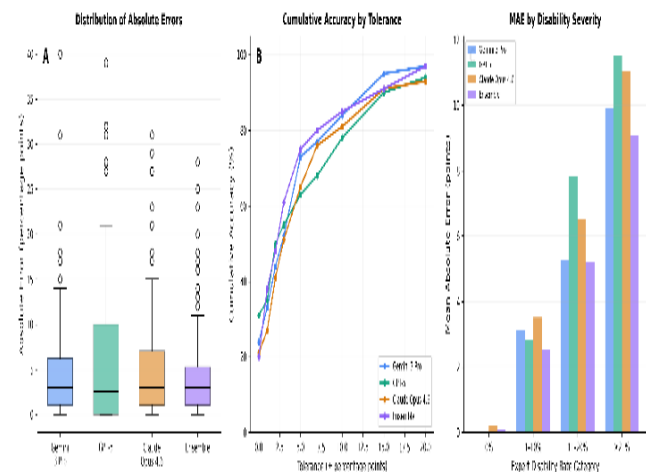


Figure 3: Error analysis of disability impairment rate predictions. (A) Box plots showing the distribution of absolute errors for each model and the ensemble. (B) Cumulative accuracy curves showing the proportion of predictions within a given tolerance threshold. (C) Mean absolute error stratified by expert disability rate category (0%, 1–10%, 11–20%, and $\geq 21\%$).

Table 4: Agreement Metrics for Temporary Work Disability Duration and Caregiver Need Duration

Temporary Work Disability Duration						
Model	n	ICC (3,1) [95% CI]	Spearman ρ	MAE (days)	Bias (days)	Exact Agreement (n)
Gemini 3 Pro	99	0.729 [0.624–0.811]	0.798	45.7	+10.5	46
GPT-5	99	0.701 [0.583–0.788]	0.765	50.5	-3.0	33
Claude Opus 4.6	99	0.750 [0.665–0.834]	0.801	43.8	+19.8	47
Three-Model Ensemble	99	0.773 [0.681–0.842]	0.825	44.2	+9.1	22
Caregiver Need Duration						
Model	n	ICC (3,1) [95% CI]	Spearman ρ	MAE (days)	Bias (days)	Exact Agreement (n)
Gemini 3 Pro	95	0.529 [0.431–0.700]	0.703	28.7	-20.5	37
GPT-5	95	0.497 [0.406–0.685]	0.688	31.4	-23.6	33
Claude Opus 4.6	95	0.585 [0.466–0.722]	0.730	27.6	-13.8	32
Three-Model Ensemble	95	0.556 [0.462–0.720]	0.737	28.8	-19.3	17

ICC, intraclass correlation coefficient; CI, confidence interval; MAE, mean absolute error.

Table 5: Agreement metrics for disability impairment rate by impairment type (single vs. multiple/Balthazard)

Single impairment (n=30)	ICC (3,1)	Spearman ρ	MAE	Bias	95% LoA	±5 pp (%)	±10 pp (%)
Gemini 3 Pro	0.970	0.853	1.9	+0.5	-6.6 to +7.5	97	97
GPT-5	0.919	0.900	2.3	+1.6	-9.6 to +12.8	93	93
Claude Opus 4.6	0.865	0.900	3.4	+2.6	-13.1 to +18.4	87	90
Ensemble	0.950	0.938	2.2	+1.6	-7.2 to +10.3	90	93
Multiple impairment (n=70)	ICC (3,1)	Spearman ρ	MAE	Bias	95% LoA	±5 pp (%)	±10 pp (%)
Gemini 3 Pro	0.826	0.745	6.5	-1.8	-20.5 to +16.9	63	79
GPT-5	0.757	0.689	7.7	-3.0	-24.3 to +18.3	50	71
Claude Opus 4.6	0.824	0.732	6.9	-1.7	-20.0 to +16.7	56	77
Ensemble	0.854	0.809	5.7	-2.2	-18.2 to +13.9	69	81
Mann-Whitney U Test Results †							
p-value		p-value		p-value		p-value	
Gemini 3 Pro	<0.001	GPT-5	<0.001	Claude Opus 4.6	<0.001	Ensemble	<0.001

ICC, intraclass correlation coefficient; MAE, mean absolute error; LoA, limits of agreement; pp, percentage points.

†Mann-Whitney U test comparing absolute errors between single and multiple impairment group

Single vs. multiple impairment analysis: A marked performance difference emerged between single and multiple impairment cases (Table 5). For single impairment cases (n=30), all models achieved excellent-to-good agreement: ICC 0.970 (Gemini 3 Pro), 0.919 (GPT-5), and 0.865 (Claude Opus 4.6), with MAE of 1.9–3.4 pp and slight positive bias. The ensemble achieved ICC 0.950 and MAE 2.2. In contrast, for multiple impairment cases (n=70), agreement was substantially lower: ICC 0.826 (Gemini 3 Pro), 0.757 (GPT-5), and 0.824 (Claude Opus 4.6), with MAE of 6.5–7.7 and negative bias. The ensemble ICC was 0.854 (MAE 5.7). The proportion of predictions within ± 5 pp dropped from 87–97% (single) to 50–63% (multiple). The between-group difference was highly significant for all models (Mann-Whitney U, all $p < 0.001$).

Secondary outcomes: Agreement for temporary work disability duration was moderate across all models, with ICC values of 0.729 (Gemini 3 Pro), 0.701 (GPT-5), and 0.750 (Claude Opus 4.6) (Table 4). MAE values ranged from 43.8 to 50.5 days. Claude Opus 4.6 demonstrated the highest positive bias (+19.8 days), indicating a tendency to overestimate work disability duration. Agreement for caregiver need duration was poor to moderate, with ICC values ranging from 0.497 (GPT-5) to 0.585 (Claude Opus 4.6) (Table 4). All three models systematically underestimated caregiver duration, with negative biases ranging from –13.8 to –23.6 days.

Discussion

To our knowledge, this is the first study to evaluate LLMs in calculating quantitative disability impairment ratings based on national regulatory frameworks. While prior studies have assessed LLM capabilities in forensic report writing (9), medical licensing examinations (3), and clinical diagnostic accuracy (7), this study addresses a distinct task: numerical computation of disability percentages requiring simultaneous clinical data extraction, regulatory table lookup, and multi-step arithmetic including the Balthazard formula. Current commercial LLMs can perform this task with good overall agreement (ICC 0.814–0.875), although accuracy varies considerably across severity levels. The near-perfect accuracy for 0% disability cases (MAE ≤ 0.2) contrasted with progressively increasing errors for higher rates (MAE 9.1–11.5 for $\geq 21\%$) has important practical implications. Lower disability rates typically involve single-system injuries assessed with a single table lookup, whereas higher rates involve multi-system injuries requiring sequential

application of the Balthazard formula—compounding errors at each step. This pattern is consistent with findings that LLM accuracy declines significantly in multi-step computational tasks (17); GPT-4o accuracy dropped from 62.7% to 43.6% under strict step-wise evaluation, revealing errors masked by lenient tolerance thresholds (17). Our findings suggest that the primary challenge for LLMs in disability assessment lies not in clinical interpretation but in the precise numerical execution of sequential regulatory computations. The subgroup analysis by impairment type further substantiates this conclusion. For single impairment cases requiring only a direct table lookup, all models achieved excellent agreement (ICC 0.865–0.970, MAE 1.9–3.4), demonstrating that LLMs can reliably extract clinical findings and match them to appropriate regulatory table entries. However, for multiple impairment cases requiring the Balthazard formula, agreement decreased substantially (ICC 0.757–0.826, MAE 6.5–7.7), with a reversal from positive to negative bias. This pattern indicates that the primary technical bottleneck is the cumulative mathematical computation inherent in the Balthazard formula, where each sequential calculation step introduces compounding errors. The highly significant difference between groups (all $p < 0.001$) confirms that computational complexity of combining multiple impairment ratings is the dominant factor driving accuracy decline. This finding carries particular clinical significance because cases with $\geq 21\%$ disability rates carry the greatest medicolegal impact—larger compensation, longer work disability periods, and greater caregiver needs. Systematic underestimation in this subgroup means uncritical reliance on LLM outputs could result in undervaluation of patients' impairments, with direct adverse consequences for their legal rights and financial compensation. The three-model ensemble outperformed every individual model across all agreement metrics (ICC 0.890, MAE 4.7), with significant improvements over Claude Opus 4.6 ($p < 0.001$) and GPT-5 ($p = 0.028$). This aligns with emerging evidence that combining outputs from multiple LLMs yields more reliable predictions than any single model. Barabucci et al. showed that aggregating differential diagnoses from four LLMs produced higher accuracy than the best individual model (18), and Yang et al. reported that ensemble frameworks consistently outperformed individual models in medical question-answering (19). Our study extends this principle to continuous quantitative estimation in a medicolegal context, suggesting that averaging

outputs from multiple AI systems may reduce errors in AI-assisted disability assessment. Performance on secondary outcomes was notably weaker. Agreement for temporary work disability duration was moderate (ICC 0.701–0.750), while caregiver need duration showed poor-to-moderate agreement (ICC 0.497–0.585), with systematic underestimation across all models. Unlike disability rate calculation, these outcomes lack an explicit tabular framework and require holistic judgment about recovery trajectories—a task current LLMs appear less equipped for. This underscores that LLM performance partially depends on structured, rule-based reference frameworks. The Turkish disability assessment landscape involves multiple regulatory frameworks: the 2019 regulation for adults (11), the 2008 regulation on work capacity loss, and the 2013 regulation on disability criteria (20). In a prior study at our institution on 548 Van disability cases, traffic accidents accounted for 91.6% and 19.5% healed without residual impairment (20). Our case distribution (traffic accidents 89%, 0% disability 9%) was designed to approximate this real-world pattern, enhancing ecological validity. Nevertheless, it must be acknowledged that simulated cases, regardless of how carefully constructed, represent a controlled testing environment that may not fully reflect the challenges of real clinical practice. Actual medicolegal cases often present with incomplete medical records, ambiguous clinical findings, comorbidities, and conflicting imaging interpretations—factors that could further challenge LLM performance. Therefore, the agreement levels observed in this study should be interpreted as an optimistic estimate, and prospective validation with real patient data is essential before any clinical implementation can be considered.

Study limitations: This study has several limitations. First, the reference standard was determined by a single forensic medicine specialist rather than a multi-rater panel. In clinical practice, disability assessment committees may reach different conclusions for the same case due to subjective interpretation of clinical findings and regulatory criteria. The absence of inter-rater reliability data for the reference standard means that the observed AI–expert disagreements may partially reflect inherent variability in expert judgment rather than AI error alone. Future studies should employ multiple independent raters to establish a more robust reference standard. Second, all clinical cases were fictional, purpose-constructed scenarios rather than actual patient

records. Although these cases were designed to reflect the clinical spectrum encountered in real forensic practice, fictional cases may not fully capture the complexity, ambiguity, and incomplete documentation characteristic of real medicolegal evaluations, which limits the external validity and generalizability of the findings. Third, the study used a zero-shot prompting approach; alternative strategies such as chain-of-thought prompting or retrieval-augmented generation may yield different results. Fourth, LLM outputs are non-deterministic, and we did not perform repeated measurements to assess intra-model consistency. Fifth, given the rapid pace of model development, the results reflect a specific point in time (February 2026) and may not generalize to future model versions. Finally, all cases were based on Turkish regulatory tables, and the findings may not be directly transferable to disability assessment systems in other jurisdictions.

Conclusion

Three leading commercial LLMs demonstrated good agreement with expert-determined disability impairment ratings across 100 standardized cases (ICC >0.80 for all models). However, a critical limitation emerged: accuracy declined substantially for disability rates $\geq 21\%$ —the cases with greatest medicolegal consequence—where MAE reached 9.1–11.5 pp and all models systematically underestimated disability rates. Subgroup analysis confirmed that this decline reflects the computational complexity of the Balthazard formula rather than clinical interpretation errors. A three-model ensemble yielded the highest agreement (ICC 0.890) and lowest error (MAE 4.7), suggesting that aggregating outputs may partially mitigate individual model errors. Despite these results, the findings do not support LLMs as independent tools in medicolegal disability assessment. The systematic underestimation in complex cases, combined with limitations inherent in simulated cases and a single-expert reference standard, necessitates considerable caution. At present, LLMs may serve only as preliminary screening aids under direct expert supervision, and their outputs must be independently verified before any medicolegal application. Further studies with real patient data, multi-rater reference standards, and prospective validation are essential before any integration into clinical practice.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest: The authors declare that they have no conflict of interest.

Informed consent: Not applicable. No real patients were involved in this study.

Author contributions: Y.E.: Conceptualization, study design, case construction, data collection, statistical analysis, writing – original draft, writing – review and editing. E.K.: Data collection, case construction, study design. M.A.: Supervision, writing – review and editing, study design, statistical analysis.

Use of artificial intelligence tools: Artificial intelligence chatbots (ChatGPT Plus, Google AI Pro, and Claude Pro) were used as the subjects of evaluation in this study. The authors take full responsibility for the accuracy and integrity of all content.

Data availability statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8
2. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. 2023;6(1):120. doi:10.1038/s41746-023-00873-0
3. Jaleel A, Aziz U, Farid G, Bashir MZ, Mirza TR, Abbas SMK, et al. Evaluating the potential and accuracy of ChatGPT-3.5 and 4.0 in medical licensing and in-training examinations: systematic review and meta-analysis. *JMIR Med Educ*. 2025;11:e68070. doi:10.2196/68070
4. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open*. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969
5. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*. 2025;333(4):319-328. doi:10.1001/jama.2024.21700
6. Lee J, Park S, Shin J, Cho B. Analyzing evaluation methods for large language models in the medical field: a scoping review. *BMC Med Inform Decis Mak*. 2024;24(1):366. doi:10.1186/s12911-024-02709-7
7. Dinc MT, Bardak AE, Bahar F, Noronha C. Comparative analysis of large language models in clinical diagnosis: performance evaluation across common and complex medical cases. *JAMIA Open*. 2025;8(3):ooaf055. doi:10.1093/jamiaopen/ooaf055
8. Dinis-Oliveira RJ, Azevedo RMS. ChatGPT in forensic sciences: a new Pandora's box with advantages and challenges to pay attention. *Forensic Sci Res*. 2023;8(4):275-279. doi:10.1093/fsr/owad039
9. Aydogan HC, Yıkar B, Balandız H, Özsoy S. Assessing ChatGPT-4's ability to generate forensic reports: a study of artificial intelligence in forensics. *Egypt J Forensic Sci*. 2025;15:30. doi:10.1186/s41935-025-00445-1
10. Sohail A, Patel OM, Choi J, Venditti JCS, Wu AJ. Benchmarking multimodal large language models for forensic science and medicine. *medRxiv*. 2025. doi:10.1101/2025.07.06.25330972
11. Erişkinler İçin Engellilik Değerlendirmesi Hakkında Yönetmelik [Regulation on Disability Assessment for Adults]. T.C. Resmî Gazete, 20 Şubat 2019, Sayı: 30692.
12. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011;64(1):96-106. doi:10.1016/j.jclinepi.2010.03.002
13. Dietterich TG. Ensemble methods in machine learning. In: Kittler J, Roli F, editors. *Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science*, vol. 1857. Berlin: Springer; 2000. p. 1-15. doi:10.1007/3-540-45014-9_1
14. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428. doi:10.1037/0033-2909.86.2.420
15. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012
16. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307-310. doi:10.1016/S0140-6736(86)90837-8

17. Wang B, Xia I, Zhang Y, Wang J, Ouyang F, Han S, et al. From scores to steps: diagnosing and improving LLM performance in evidence-based medical calculations. arXiv. 2025. doi:10.48550/arXiv.2509.16584
18. Barabucci G, Shia V, Chu ES, Gong JJ, Goodman KE, Meltzer AC, et al. Combining multiple large language models improves diagnostic accuracy. NEJM AI. 2024;1(11):AIcs2400502. doi:10.1056/AIcs2400502
19. Yang H, Li M, Zhou H, Xiao Y, Fang Q, Zhou S, et al. Large language model synergy for ensemble learning in medical question answering: design and evaluation study. J Med Internet Res. 2025;27:e70080. doi:10.2196/70080
20. Hekimoğlu Y, Gümüş O, Kartal E, Etli Y, Demir U, Aşırdizer M. The evaluation of relationship between disability rates and age and gender. Van Med J. 2017;24(3):173-181. doi:10.5505/vtd.2017.53244