Research Article

Let the History Speak: Zero-Shot LLMs for Diagnosing Vestibular Disorders

Chongkai Lu[1], Ruiqi Zhang[1], Fangzhou Yu[1], Huawei Li[1,2,3,4,5], Peixia Wu[1,6*]

[1] *Department of Otorhinolaryngology, Eye & ENT Hospital, Fudan University, Shanghai, China*
[2] *State Key Laboratory of Medical Neurobiology and Ministry of Education Frontiers Center for Brain Science, Fudan University, Shanghai, China*
[3] *National Health Commission Key Laboratory of Hearing Medicine, Fudan University, Shanghai, China*
[4] *Institutes of Brain Science and the Collaborative Innovation Center for Brain Science, Fudan University, Shanghai, China*
[5] *Institutes of Biomedical Sciences, Fudan University, Shanghai, China*
[6] *Vertigo and Balance Disorder Center, Eye & ENT Hospital, Fudan University, Shanghai, China*
[*] *Corresponding author. E-mail: 13524844562@163.com*

ABSTRACT

Background: Vestibular disorders are common yet diagnostically challenging in first-line and specialist settings, and delays or misclassification can alter management and outcomes. Structured symptom questionnaires and supervised machine learning (ML) have shown promise for triage, while recent large language models (LLMs) may reason over clinical descriptions without task-specific training.

Objective: To evaluate zero-shot LLMs for five-class vestibular diagnosis from an electronic questionnaire, characterize error patterns across disorders, and compare the best-performing LLM with a trained gradient-boosted tree (LightGBM, LGBM).

Methods: We used a seven-center prospective cohort with an electronic 23-item questionnaire and guideline-based reference diagnoses by experienced ENT specialists. The prediction task was a five-class classification among benign paroxysmal positional vertigo (BPPV), vestibular migraine (VM), Meniere disease (MD), sudden sensorineural hearing loss with vestibular dysfunction (SSNHL-V), and an aggregated "Others" category of individually rare vestibular conditions. After prespecified exclusions, 1,025 single-definite cases were analyzed; 912/113 patients formed the train/test split for a LightGBM baseline. Three LLMs (DeepSeek-R1, DeepSeek-V3, Doubao-1.6-thinking) were evaluated zero-shot on all 1,025 cases. We report Top-k, MRR, and NDCG@5 overall; one-vs-rest sensitivity, specificity, and accuracy per disorder (macro-averaged where applicable); 95% CIs via 1,000-patient bootstrap; paired bootstrap for model differences; and McNemar's test for accuracy on the shared test set.

Results: All LLMs outperformed a prevalence prior baseline (Top-1 38.6%). V3 and Doubao achieved Top-1 ≈ 65% and Top-3 ⩾ 91%, with MRR ≈ 0.79–0.80. Disorder-wise, BPPV was reliably detected; vestibular migraine remained hardest; sensitivity–specificity trade-offs were model- and disorder-dependent. On the 113-case test set, LGBM slightly exceeded V3 on sensitivity (0.722 vs. 0.632), specificity (0.941 vs. 0.926), and accuracy (0.770 vs. 0.742), with no significant accuracy difference (McNemar $p = 0.690$). Findings support LLMs as a zero-shot front end that narrows diagnostic search space while approaching a specialized model's performance.

KEYWORDS
vestibular disorder, vertigo, large language model, medical history

1 Introduction

Dizziness and vertigo are common, high-impact presentations in primary care, emergency, and specialty clinics, yet early classification remains difficult because brief positional vertigo, migrainous features, fluctuating auditory symptoms, and acute audiovestibular events often overlap [1, 2]. Guideline frameworks (e.g., AAO–HNS

Article history: Received Aug 15, 2025; Accepted Aug 26, 2025

for BPPV; Bárány Society criteria for peripheral vestibular disorders) provide definitional clarity, but their application at scale still hinges on complete, structured history-taking and consistent clinical synthesis [3, 4]. Delays or misclassification can alter downstream testing and treatment, underscoring the need for reliable, scalable decision support grounded in patient-reported history [5, 6].

Electronic, structured questionnaires address part of this need by standardizing symptom capture and creating analyzable inputs for decision support [7, 8]. Supervised machine learning (ML) models trained on such instruments have reported encouraging performance and can be tuned for calibration and specificity; however, they require labeled training data, re-training under distribution shift, and feature maintenance across sites [9, 10]. Meanwhile, large language models (LLMs) have rapidly advanced in zero-/few-shot reasoning over clinical text and semistructured inputs, raising the prospect of flexible, promptable diagnostic support from patient-reported information [11]. What is still limited are rigorous, multicenter, head-to-head evaluations that place zero-shot LLMs and task-specific supervised models on the *same* prospective cohort, quantify uncertainty, and map error structure at the level of individual vestibular disorders [11].

To address these gaps, we studied a seven-center prospective cohort using an electronic 23-item questionnaire to perform a *five-class* vestibular diagnosis task (BPPV, vestibular migraine, Meniere disease, SSNHL-V, and a composite "Others"). We evaluated three contemporary LLMs in a zero-shot setting and compared the top LLM with a LightGBM (LGBM) model trained on the same instrument. Evaluation combined overall ranking metrics (Top-k, MRR, NDCG@5) with per-disorder one-vs-rest sensitivity, specificity, and accuracy; uncertainty was quantified with patient-level bootstrap and paired bootstrap for model differences, and McNemar's test was used for accuracy on the shared external test set [12, 13]. At a high level, LLMs delivered strong ranking quality (Top-1 around two-thirds and Top-3 above ninety percent) and exhibited clinically coherent disorder-wise profiles; the best LLM achieved accuracy comparable to the trained LGBM on the external test subset, with a non-significant difference. Error anatomy highlighted reliable recognition of BPPV, persistent difficulty for vestibular migraine, a sensitivity–specificity split for Meniere disease, and high specificity with favorable sensitivity for SSNHL-V.

Our contributions are threefold. First, we provide a prospective, multicenter, head-to-head assessment of zero-shot LLMs against a task-specific supervised comparator on a standardized, five-class vestibular instrument—closing a key evidence gap for history-driven decision support [14]. Second, we pair aggregate metrics with rigorous uncertainty quantification and a clinically interpretable error analysis (including confusion structures) that identifies where prompt design or guardrails can curb systematic confusions. Third, we outline an *LLM-centered* workflow that is immediately practical with fixed questionnaires and, crucially, extends to conversational intake: the same models can conduct adaptive patient interviews, reconcile inconsistencies, and generate uncertainty-aware differentials for clinician review — capabilities with potential to improve diagnostic fidelity, patient acceptance, and clinic throughput when deployed with appropriate safety, calibration, and governance measures [15].

2 Methods

2.1 Cohort and Data Collection

This work uses the prospective, multicenter cohort established in seven tertiary ENT/vertigo clinics (August 2019–March 2021), namely: ENT and vertigo clinics of Eye & ENT Hospital of Fudan University; The Second Hospital of Anhui Medical University; The First Affiliated Hospital of Xiamen University; Shengjing Hospital of China Medical University; Shanghai Pudong Hospital; Shenzhen Second People's Hospital; and The First Affiliated Hospital of Chongqing Medical University. At the first specialist visit, eligible patients completed an *electronic* diagnostic questionnaire on a tablet or smartphone after informed consent; for those unable to complete it independently, trained staff read the questions aloud and recorded responses. Routine clinical care and follow-up proceeded without protocol interference. Reference diagnoses were assigned by ENT specialists (> 5 years of experience) who were blinded to questionnaire responses and applied guideline-based criteria (AAO-HNS for BPPV and Bárány Society criteria for other vestibular disorders).

In total, 1,760 patients were approached and 1,693 enrolled after consent (96.2% response; 67 declined). Of the enrolled, 1,041 received a single, final diagnosis within the two-month follow-up window. For evaluation reliability, we excluded cases with multiple diagnoses (n = 14), only probable diagnoses (n = 145), undetermined diagnoses (n = 493), and an additional 16 records with contradictory entries identified during pre-analysis quality control, yielding 1,025 single-definite cases for analysis (Figure 1). For traditional machine learning, 912 cases were used for training and 113 for testing (the held-out external test set from the published cohort). LLMs required no training and were evaluated on all 1,025 cases; for head-to-head comparison with traditional models, metrics were computed on the shared test set (n = 113). Key demographics and case-mix are summarized in Table 1: median age was 54 years (IQR 41–65), 56.5% were female, and BPPV was the most frequent single

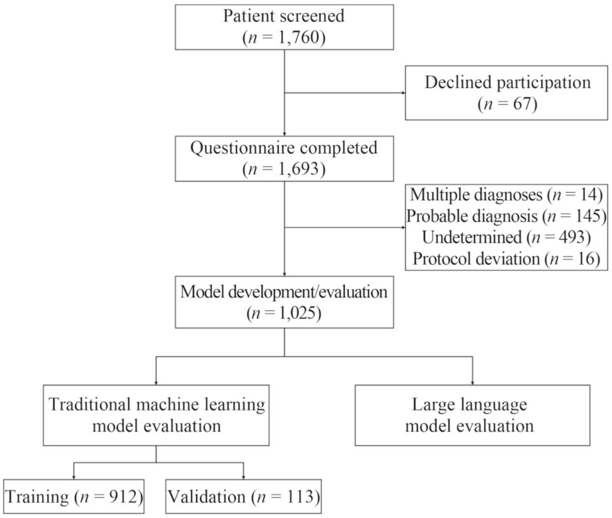

Fig. 1 Participant flow for the seven-center cohort, enrollment, exclusions, and analysis sets.

Table 1 Demographic and clinical characteristics of the study population

Characteristic	Value
Total participants, *n*	1025
Age (year), median (IQR)	54 (41–65)
Sex, *n* (%)	
Female	579 (56.5)
Male	397 (38.7)
Diagnoses, *n* (%)	
Benign paroxysmal positional vertigo	396 (38.6)
Vestibular migraine	206 (20.1)
Ménière disease	198 (19.3)
Sudden sensorineural hearing loss with vestibular dysfunction	157 (15.3)
Others	68 (6.6)

diagnosis (38.6%), followed by vestibular migraine (20.1%), Meniere disease (19.3%), and sudden sensorineural hearing loss with vestibular dysfunction (15.3%).

2.2 Questionnaire: design, administration, and content

The diagnostic questionnaire was built through a three-stage iterative process —focus/panel meetings (drafting disorder features), patient cognitive interviews (simplifying wording and pruning items), and an expert panel review (reordering/merging items). The final instrument comprised 23 items with branching logic and was administered electronically as above. Content covered: symptom character; attack frequency/duration and time since first onset; laterality and dynamics of hearing loss; tinnitus/aural-fullness/earache around attacks; headache features and family history; photophobia/phonophobia; unsteadiness and worsening with standing/walking; falls/consciousness/incontinence during attacks; common triggers (positional change, Valsalva/sound/pressure, visually complex scenes, foods/odors, fatigue/insomnia/anger); cervicogenic clues (upper-limb numbness/neck pain); prodromal infections; and otologic/trauma history.

For modeling and reporting, diagnostic categories followed the published work [10]: BPPV, vestibular migraine, Menière disease, sudden sensorineural hearing loss with vestibular dysfunction (SSNHL-V), and an “Others” bin for individually rare conditions (e.g., vestibular neuritis, PPPD, bilateral vestibulopathy, psychogenic dizziness, delayed endolymphatic hydrops, vestibular paroxysmia, cervicogenic vertigo, acoustic neuroma, presbyvestibulopathy, light cupula, Ramsay–Hunt syndrome, labyrinthine fistula, and superior canal dehiscence).

2.3 Evaluation metrics and uncertainty quantification

All point estimates are computed at the patient level. Unless stated otherwise, large-language model (LLM) ranking metrics are calculated on the full analysis set (n = 1,025), whereas the head-to-head comparison between the best LLM and the LightGBM (LGBM) baseline uses the shared external test set (n = 113). *Classification metrics.* Sensitivity and specificity are computed for each diagnosis in a one-vs-rest manner and then averaged with equal weight across the five classes (macro average). Overall accuracy is the proportion of correct top-1 predictions. *Ranking metrics.* Top-k accuracy counts a case as correct if the reference label appears within the model’s top k. Mean reciprocal rank (MRR) averages the inverse of the position of the correct label. NDCG@5 rewards a correct label near the top of the list and is normalized to 1 for an ideal ranking. *Uncertainty and paired testing.* All error bars shown in figures are 95% confidence intervals from 1,000 patient-level bootstrap samples. For paired model comparisons on the same patients, we use a paired bootstrap to form CIs for metric differences; for accuracy, we additionally report McNemar’s test. Unless noted, p-values are two-sided and no multiple-comparison adjustment is applied.

3 Results

3.1 Large Language Model Performance Evaluation

Across the five-class diagnostic ranking task, all three LLMs substantially outperformed a prevalence-based prior baseline (Table 2; Figure 2). Against the baseline (Top-1 38.6%, MRR 0.603, NDCG@5 0.701), *DeepSeek-R1* improved Top-1 by +15.4 percentage points (54.0%), *DeepSeek-V3* by +26.6 points (65.2%), and *Doubao-1.6-thinking* by +27.0 points (65.6%). Gains were consistent when allowing more candidates: Top-3 reached 88.5%

Table 2 Model Performance on Vestibular Disease Classification

Model	Top-1% (95% CI)	MRR (95% CI)	NDCG@5 (95% CI)	*p*
Baseline	38.6	0.603	0.701	–
DeepSeek-R1	54.0 (51.0, 57.2)	0.719 (0.700, 0.739)	0.790 (0.775, 0.805)	< 0.001
DeepSeek-V3	65.2 (62.2, 68.1)	0.789 (0.770, 0.807)	0.842 (0.828, 0.856)	< 0.001
Doubao-1.6-think	65.6 (62.6, 68.6)	0.795 (0.778, 0.814)	0.847 (0.834, 0.861)	< 0.001

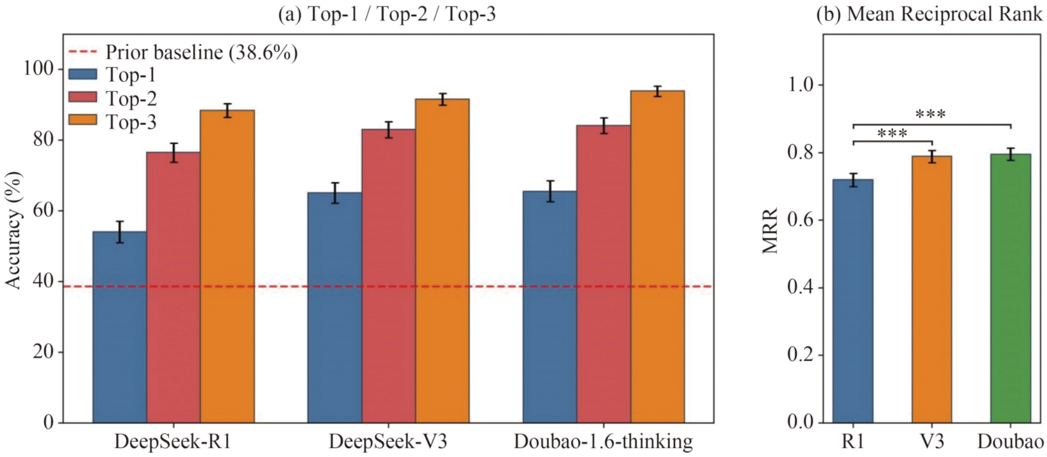

Fig. 2 Overall LLM ranking performance on the five-class diagnostic task. Left: Top-1/Top-2/Top-3 accuracy, with the dashed line showing the prior baseline; all models exceed baseline, and DeepSeek-V3 and Doubao form a clear top tier while R1 lags. Right: Mean Reciprocal Rank; R1 is significantly lower than both V3 and Doubao (***), whereas V3 and Doubao are essentially indistinguishable. Error bars denote 95% CIs.

(R1), 91.7% (V3), and 94.0% (Doubao). These absolute improvements translate into stronger ranking quality, with mean reciprocal rank (MRR) of 0.719, 0.789, and 0.795, respectively, and NDCG@5 of 0.790, 0.842, and 0.846.

Model ordering was stable across metrics: DeepSeek-R1 formed the lower tier, while DeepSeek-V3 and Doubao-1.6-thinking clustered at the top with very similar point estimates. Notably, the leading model depends on the metric—Doubao is marginally higher on Top-1 (65.6% vs. 65.2% for V3) and ranking metrics (MRR 0.795 vs. 0.789; NDCG@5 0.846 vs. 0.842 in Table 2). Uncertainty estimates (95% CIs) were obtained by bootstrapping over patients and are narrow enough to support the above ordering. Pairwise MRR tests confirm that R1 is significantly below both V3 and Doubao ($p = 8.4\times10^{-14}$ and 3.2×10^{-15}), whereas V3 and Doubao are statistically indistinguishable on MRR ($p = 0.41$). Taken together, these results suggest that, on this structured clinical questionnaire, current frontier LLMs deliver robust ranking performance: a two-thirds Top-1 hit rate without task-specific training, and high Top-3 coverage exceeding 90%, which could meaningfully reduce the downstream diagnostic search space for clinicians.

3.2 Diagnostic Error Pattern Analysis

Class-wise performance shows distinct difficulty profiles across disorders (Figure 3). BPPV is reliably recognized (sensitivities: R1 86.6%, V3 83.6%, Doubao 82.6%), but R1's higher sensitivity is accompanied by substantially lower specificity (64.4%) compared with V3 (82.2%) and Doubao (84.7%), both significantly higher; Doubao's specificity also exceeds V3's slightly. Vestibular migraine (VM) remains the most challenging named disorder: V3 improves sensitivity over R1 (52.9% vs 43.2%; significant), while Doubao yields the best specificity (89.1%), significantly above both R1 and V3, at a sensitivity comparable to V3 (49.0%; not significant). For Meniere disease, V3 is sensitivity-oriented (71.2%), significantly exceeding both R1 and Doubao, whereas Doubao is specificity-oriented (94.8%), significantly higher than V3 and R1 with intermediate sensitivity (60.6%). In SSNHL-V, the main separation is sensitivity: Doubao (60.5%) > V3 (51.0%; significant) ≫ R1 (24.8%; highly significant), while all three maintain similarly high specificity (around 98%; no significant differences). Finally, in the heterogeneous Others category, Doubao markedly increases sensitivity (42.6% vs. 11.8% for R1 and 10.3% for V3; both highly significant) at the cost of lower specificity (88.7% vs. around 97%–98%; both highly significant). Overall, these patterns indicate sensitivity–specificity trade-offs: V3 favors sensitivity in Meniere disease, Doubao favors specificity in VM and sensitivity in SSNHL-V/Others, and R1 tends to over-call BPPV.

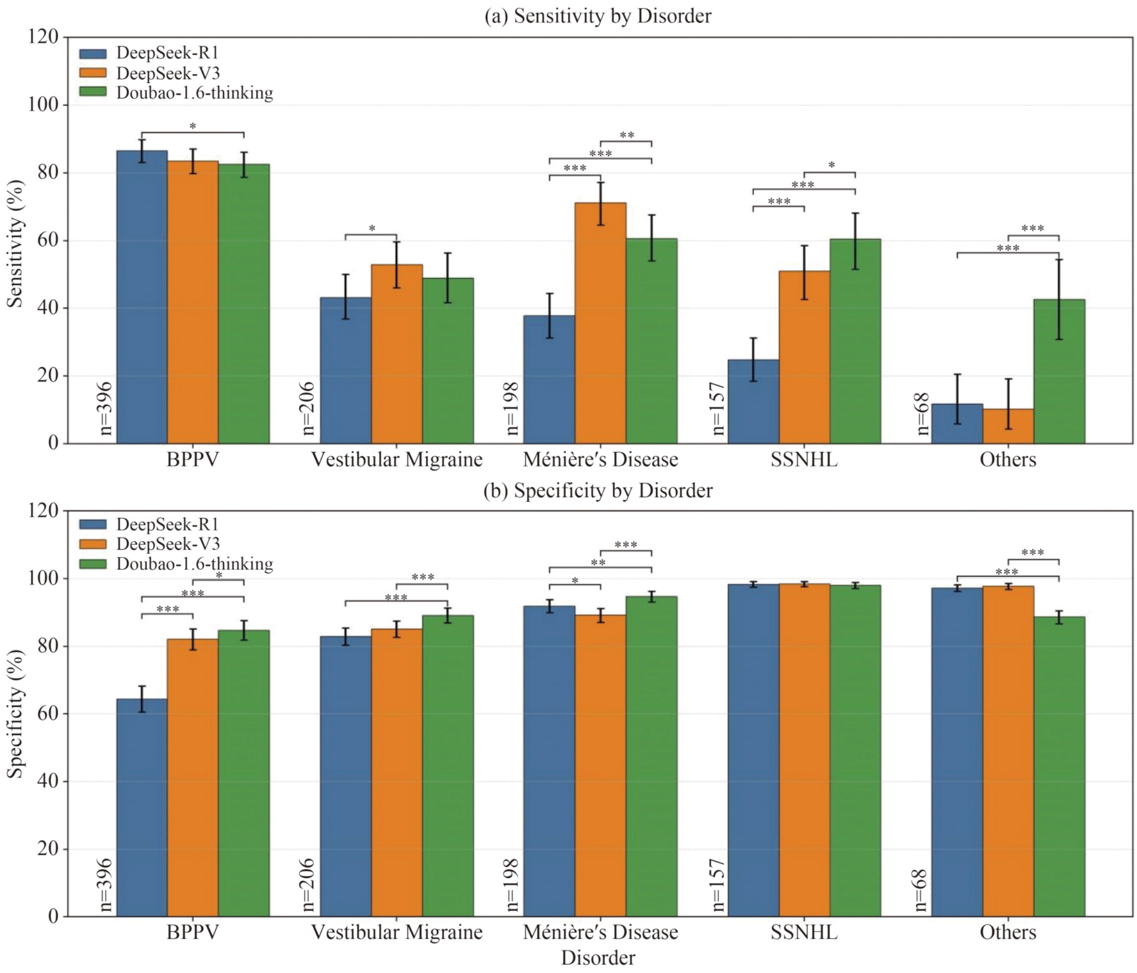

Fig. 3 Per-disorder sensitivity and specificity with pairwise significance annotations. Error bars denote 95% CIs.

The confusion matrices and diagnostic-flow visualization (Figures 4 and 5) reveal three recurring error channels. First, a prominent VM↔BPPV axis: a large share of VM is labeled as BPPV (V3 27.7%; Doubao 25.7%), with a smaller but clear BPPV→VM spillover (V3 10.4%). Second, SSNHL-V→Meniere disease is frequent (V3 19.7%, R1 23.6%), and is reduced by Doubao (9.6%), indicating that acute audiovestibular cues are sometimes attributed to endolymphatic pathology when questionnaire signals are ambiguous. Third, the composite Others category disperses broadly into common peripheral disorders; Doubao's higher sensitivity increases on-diagonal hits (42.6%) but also raises false positives, consistent with its lower specificity.

Taken together, the per-class metrics and confusion structure are clinically plausible: BPPV is a high-signal target detected well by all models (with R1 overprediction), VM remains difficult due to heterogeneous symptom constellations, Meniere disease exhibits a clear sensitivity–specificity split (V3 for case-finding, Doubao for rule-in), and SSNHL-V benefits from models that better leverage acute auditory cues (Doubao > V3 $\gg$ R1 in sensitivity with uniformly high specificity). These insights indicate where LLMs already narrow the diagnostic search space and where domain-aware prompting could further curb systematic errors.

3.3 Comparison with Traditional Machine Learning Approaches

3.3.1 Head-to-head performance on the held-out test set

On the shared external test set (n = 113), the traditional gradient-boosted trees model (LGBM; trained on 912 cases) achieved slightly higher point estimates than the zero-shot LLM (DeepSeekV3) on all three scalar metrics (Figure 6, left panel). Sensitivity was 0.722 for LGBM versus 0.632 for V3; specificity was high for both (0.941 vs. 0.926); and overall accuracy was 0.770 for LGBM versus 0.742 for V3. Error bars (patient-level bootstrap 95% CIs) overlap broadly, and a paired McNemar test for accuracy yielded p = 0.690 (annotated in the figure), indicating no statistically significant difference. The win–loss breakdown further illustrates the small margin (Figure 6, right panel): among 113 patients, the models agreed on 78% of cases (both correct 65%, both wrong

13%); in the remaining 22%, V3 outperformed LGBM in 10% (11/113) and underperformed in 12% (14/113), corresponding to a ΔAccuracy of −2.7 percentage points in favor of LGBM.

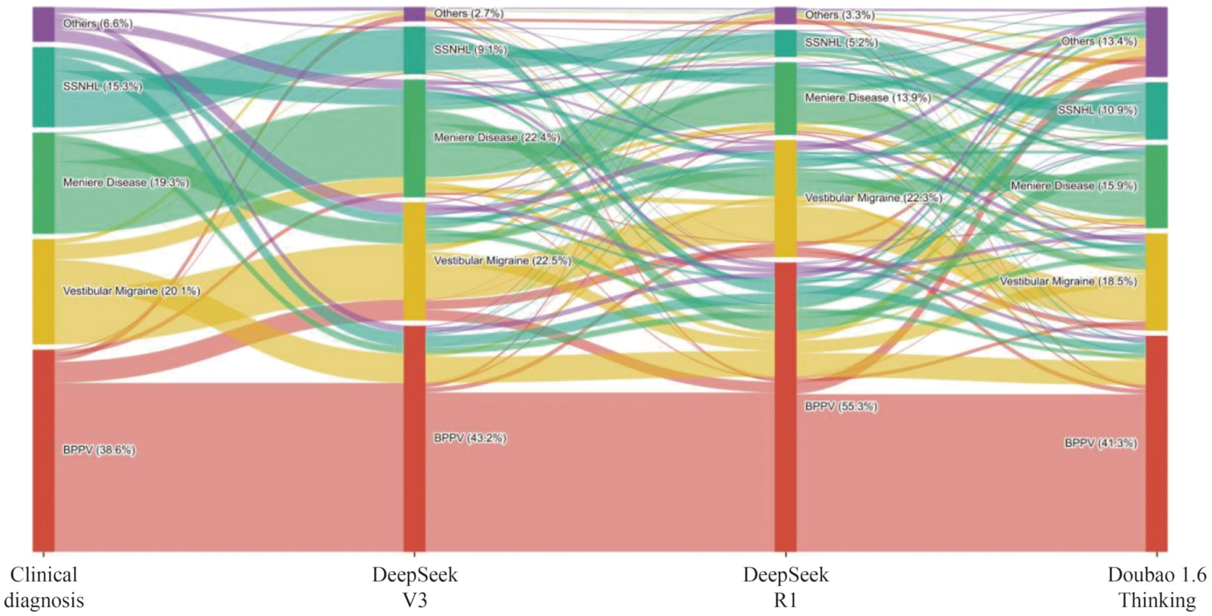

Fig. 4 Alluvial diagram of flows from clinical diagnoses to model predictions across three LLMs (DeepSeek V3, DeepSeek R1, Doubao 1.6). Bands are colored by the clinical diagnosis. A comparatively large stream is observed from clinically diagnosed Vestibular Migraine (VM) to model-predicted BPPV, with a smaller reverse BPPV→VM stream.

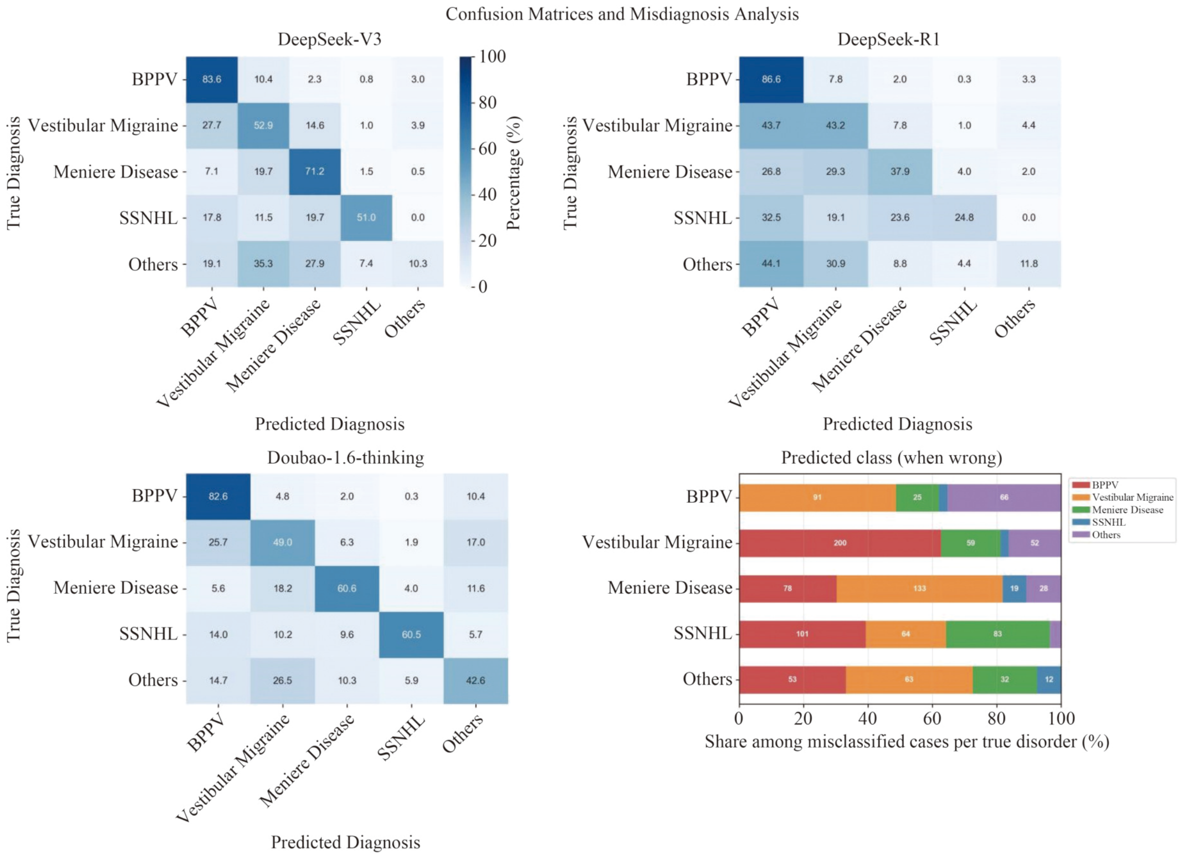

Fig. 5 Confusion matrices (row-normalized) for each model and stacked bars summarizing the distribution of predicted labels among misclassified cases.

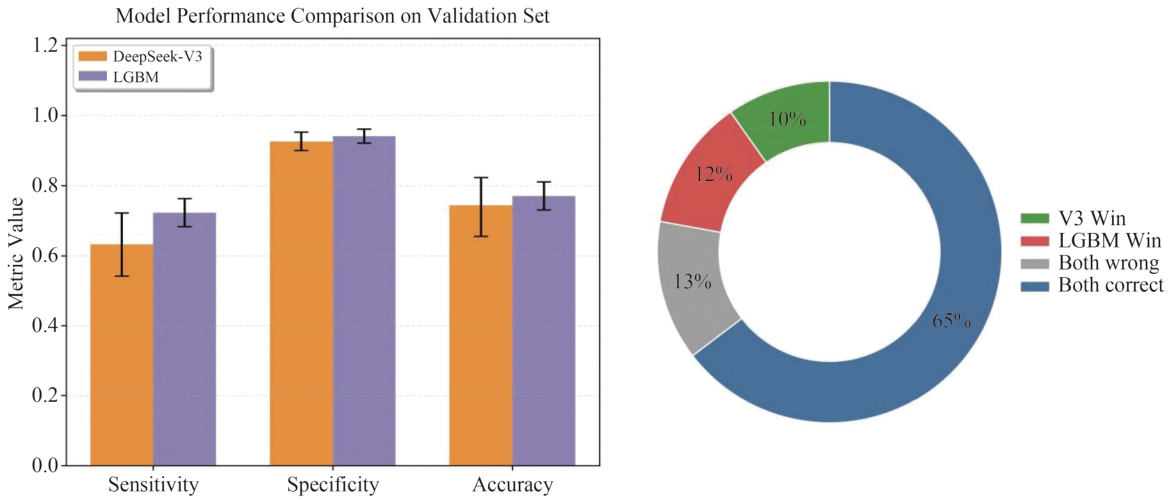

Fig. 6 Head-to-head comparison of DeepSeek-V3 (zero-shot LLM) vs. LightGBM (supervised) on the shared test set (n = 113). Left: LGBM attains slightly higher sensitivity, specificity, and accuracy, but the paired McNemar test indicates no significant accuracy difference. Right: The models agree on most cases; among disagreements, V3's wins and losses are of similar magnitude, yielding only a small overall edge for LGBM.

3.3.2 Disorder-specific trade-offs and diagnostic weight

Class-wise estimates (Table 3) help explain the near-tie overall. LGBM shows consistently higher *specificity* for common disorders (e.g., BPPV 0.94 vs. 0.82; Meniere disease 0.97 vs. 0.94) while maintaining comparable sensitivity, yielding larger positive likelihood ratios for rule-in decisions (BPPV + LR 13.86; Meniere disease + LR 21.19). In contrast, the LLM displays standout performance for SSNHL-V with vestibular symptoms: specificity reached 1.00 with high sensitivity (0.89), giving an infinite +LR and a small –LR (0.11), properties desirable for flagging this time-sensitive condition. For vestibular migraine, both methods have similar specificity (0.88) with modest sensitivities (0.48–0.57), consistent with the heterogeneous symptom profiles observed earlier. Net accuracy therefore favors LGBM in higher-prevalence categories (BPPV, Meniere disease, Others) but favors V3 in SSNHL-V; after prevalence weighting on the test set, these effects largely cancel, producing the small, non-significant aggregate gap.

3.3.3 Representative case comparisons

The case vignettes in Table 4 illustrate complementary error tendencies. Case 1 (BPPV) shows the LLM correctly prioritizing classic positional triggers despite distracting contextual details, whereas LGBM favored

Table 3 Predictive ability comparison between LLM and traditional ML models.

	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)
Benign paroxysmal positional vertigo			
DeepSeek-V3	0.92 (0.83-0.98)	0.83 (0.73-0.91)	0.87 (0.80-0.93)
LGBM 0.88	0.88 (0.78-0.96)	0.94 (0.87-0.99)	0.91 (0.86-0.96)
Vestibular migraine			
DeepSeek-V3	0.48 (0.26-0.70)	0.88 (0.81-0.94)	0.81 (0.73-0.88)
LGBM	0.57 (0.33-0.79)	0.88 (0.81-0.95)	0.82 (0.75-0.88)
Meniere disease			
DeepSeek-V3	0.73 (0.55-0.89)	0.94 (0.89-0.99)	0.89 (0.83-0.95)
LGBM	0.73 (0.56-0.90)	0.97 (0.92-1.00)	0.91 (0.86-0.96)
Sudden sensorineural	**hearing loss with vest**	**ibular dysfunction**	
DeepSeek-V3	0.89 (0.62-1.00)	1.00 (1.00-1.00)	0.99 (0.97-1.00)
LGBM	1.00 (1.00-1.00)	0.94 (0.89-0.98)	0.95 (0.90-0.98)
Others			
DeepSeek-V3	0.14 (0.00-0.50)	0.98 (0.95-1.00)	0.93 (0.88-0.97)
LGBM	0.43 (0.00-0.83)	0.98 (0.95-1.00)	0.95 (0.90-0.98)

Table 4 Representative case studies comparing LLM and machine learning model predictions

Case 1: A 32-year-old man presented within the past month with a single episode of spinning vertigo lasting no more than 1–2 minutes. The attack was clearly precipitated by positional change—lying down, turning in bed, or rising quickly—and was not triggered by Valsalva maneuvers or loud sounds; visually complex environments did not provoke symptoms. He noted that fatigue, poor sleep, or emotional stress preceded the spell, and standing or ambulation worsened symptoms during the episode. He denied hearing loss, tinnitus, and aural fullness, as well as upper-limb numbness or pain. There was no history of otorrhea, chronic otitis media, or ear surgery, and no recent head or neck trauma.		
Clinical Diagnosis: BPPV	LLM Prediction: 1.BPPV, 2.Vestibular Migraine	ML Model Prediction: 1.Vestibular Migraine, 2. BPPV
Case 2: A 38-year-old woman presented with a 5-month history of recurrent spinning vertigo. Individual episodes lasted no more than 1–2 minutes and occurred with variable frequency. Attacks were commonly provoked by positional change—lying down, turning in bed, or rising quickly—and could also be triggered or exacerbated by visual motion such as moving scenes or complex patterns; they were not brought on by Valsalva maneuvers or loud sounds, and she noted no dietary precipitants. Episodes often followed fatigue, poor sleep, or emotional stress. During spells she experienced gait unsteadiness, with standing and ambulation worsening symptoms; between attacks she was essentially stable. Otologic symptoms included a 5-month history of fluctuating right-sided hearing loss that tended to worsen during attacks and then partially recover, bilateral tinnitus without peri-attack change, and bilateral aural fullness; she denied otalgia. The illness was preceded by a period of upper-respiratory or gastrointestinal symptoms (fever, cold, vomiting, or diarrhea). She had no history of otorrhea, chronic otitis media, or ear surgery, and no recent head or neck trauma or operations. She denied numbness or pain in the upper limbs.		
Clinical Diagnosis: BPPV	LLM Prediction: 1. Meniere Disease, 2. BPPV	ML Model Prediction: 1. BPPV, 2. Meniere Disease
Case 3: A 30-year-old woman reported recurrent, brief attacks of spinning vertigo, each lasting no more than 1–2 minutes and occurring only every few months to years. Episodes were often precipitated by positional change (lying down, turning, or rising), with no provocation by Valsalva maneuvers, loud sounds, visual motion, or specific foods. Fatigue, poor sleep, or emotional stress frequently preceded attacks; standing or walking did not worsen symptoms. Otologic history was notable for left-sided hearing loss of sudden onset persisting for over a year, accompanied by ipsilateral tinnitus that typically intensified before attacks and eased afterward, and left-sided aural fullness. She denied upper-limb numbness or pain and had no prior otologic surgery or chronic ear disease, nor any recent head or neck trauma.		
Clinical Diagnosis: Disease	Meniere LLM Prediction: 1. BPPV, 2. Vestibular Migraine	ML Model Prediction: 1.SSNHL-V, 2. Meniere Disease

vestibular migraine—suggesting the LLM's strength in leveraging long-range semantic cues. Case 2 (BPPV) shows the converse: LGBM correctly identifies BPPV while the LLM overweights fluctuating auditory complaints and ranks Meniere disease first, reflecting its sensitivity to otologic descriptors when positional information is present but not dominant. Case 3 (labeled Meniere disease) highlights a failure mode shared by both models when brief, positional vertigo co-occurs with chronic unilateral auditory symptoms; neither model resolved the mixed signal reliably.

Overall the broader message is that a general solution (LLM, zero-shot) has reached parity adjacent performance with a *specialized* solution (LGBM trained on domain cases). Given the LLM's advantages in deployment friction (no training), interactive explainability, and rapid adaptation through prompting, these results argue for a combined model–assistant paradigm in which the LLM front-ends patient interaction and hypothesis generation, while a lightweight supervised model provides calibration and high-specificity checks for common peripheral disorders.

4 Conclusion

In this seven-center prospective evaluation of a *five-class* vestibular diagnosis task, contemporary large language models (LLMs) used in a zero-shot manner on a structured 23-item questionnaire achieved competitive—and practically useful—performance. LLMs consistently surpassed a prevalence baseline and delivered strong ranking quality (Top-1 ~ 65%, Top-3 > 90%, MRR/NDCG@5 in the 0.79 - 0.85 range), narrowing the diagnostic search space without task-specific training. Against a purpose-trained LightGBM comparator included solely as a reference, the best LLM showed parity-adjacent accuracy on an external test set (0.742 vs. 0.770; McNemar p = 0.690), underscoring that modern, general-purpose LLMs can match specialized classifiers while offering advantages in promptability, deployment simplicity, and rationale generation.

Disorder-wise patterns were clinically coherent and actionable. LLMs reliably recognized high signal BPPV, demonstrated uniformly high specificity for SSNHL-V while maintaining the most favorable sensitivity profile among models, and revealed a sensitivity–specificity split for Meniere disease (case-finding versus rule-in emphasis). Vestibular migraine remained the most challenging entity, with a prominent VM↔BPPV confusion axis; the heterogeneous "Others" category highlighted a sensitivity gain at a manageable specificity trade-off. These findings indicate that an *LLM-centered* workflow is already viable: LLMs can front-end history-based differential generation with transparent reasoning, with a lightweight supervised checker optionally layered for calibration or rule-in specificity where clinically warranted.

Although our evaluation standardized inputs via a fixed questionnaire, LLMs are not constrained to forms. The same models can conduct adaptive, conversational intake—asking clarifying follow-ups, probing timing and triggers, reconciling inconsistencies, and summarizing the differential with uncertainty-aware guidance. Such interactive acquisition is poised to further improve diagnostic fidelity and, moreover, to enhance patient acceptance and clinic throughput through natural, "human-like" dialogue. Real-world deployment should pair this capability with guardrails (calibration, abstention/escalation rules, and auditing) and prospective monitoring, but the central message stands: LLMs are ready to serve as the primary engine for history-driven vestibular diagnosis, with traditional models retained as comparators to sharpen specificity when needed.

Acknowledgments

We thank all participating patients and the clinical teams at the seven collaborating sites for their support: Eye & ENT Hospital of Fudan University; The Second Hospital of Anhui Medical University; The First Affiliated Hospital of Xiamen University; Shengjing Hospital of China Medical University; Shanghai Pudong Hospital; Shenzhen Second People's Hospital; and The First Affiliated Hospital of Chongqing Medical University. We are grateful to the research nurses and coordinators who assisted with questionnaire administration and data quality control across centers.

Ethical statement

The study was conducted in accordance with the Declaration of Helsinki and relevant local regulations. The protocol received ethics approval from the Ethics Committee of the Eye & ENT Hospital of Fudan University (Shanghai, China) via expedited review. All participants (or their legally authorized representatives, where applicable) provided written informed consent before any study procedures. Patient privacy was protected through the de-identification of all records prior to analysis.

Conflicts of interest

Professor Huawei Li, Editor-in-Chief of ENT Discovery, and Peixia Wu, Executive Editor-in-Chief of ENT Discovery, were not involved in the peer-review process or in any editorial decisions regarding this manuscript. The peer-review process was handled independently by other qualified editors to minimize potential bias. All other authors declare that they have no conflicts of interest.

Funding information

This work was supported by the National Key Research and Development Program of China (2023YFC2508000).

Data availability statement

The dataset contains individually identifiable clinical information and is not publicly available due to privacy and ethics restrictions. De-identified data underlying the findings, along with analysis code, can be shared upon reasonable request to the corresponding author and after execution of a data use agreement and institutional approvals.

Author contributions

Using the CRediT taxonomy, P.W. and H.L. conceived the study, provided resources, supervised the work, and oversaw project administration; C.L., R.Z., and F.Y. developed the methodology, with C.L. and R.Z. implementing the software, performing the formal analyses, and preparing the visualizations; C.L., R.Z., and F.Y., together with site investigators at the seven hospitals, conducted the investigation and curated the data; H.L. secured funding. C.L., R.Z., and F.Y. drafted the manuscript, and all authors reviewed, edited, and approved the final version.

References

1. Kerber KA, Callaghan BC, Telian SA, Meurer WJ, Skolarus LE, Carender W, et al. Dizziness symptom type prevalence and overlap: a US nationally representative survey. *Am J Med*. 2017, 130(12): 1465.e1–1465.e9.
2. Grill E, Strupp M, Müller M, Jahn K. Health services utilization of patients with vertigo in primary care: a retrospective cohort study. *J Neurol*. 2014, 261: 1492–1498.
3. Bhattacharyya N, Gubbels SP, Schwartz SR, Edlow JA, El-Kashlan H, Fife T, et al. Clinical practice guideline: benign paroxysmal positional vertigo (update). *Otolaryngol Head Neck Surg*. 2017, 156(3): S1–S47.
4. Lopez-Escamez JA, Carey J, Chung WH, Goebel JA, Magnusson M, Mandalà M, et al. Diagnostic criteria for Menière's disease. *J Vestib Res*. 2015, 25(1): 1–7.
5. Kerber KA, Newman-Toker DE. Misdiagnosing dizzy patients: common pitfalls in clinical practice. *Neurol Clin*. 2015, 33(3): 565–575.
6. Tehrani ASS, Coughlan D, Hsieh YH, Mantokoudis G, Korley FK, Kerber KA, et al. Rising annual costs of dizziness presentations to US emergency departments. *Acad Emerg Med*. 2013, 20(7): 689–696.
7. Roland LT, Kallogjeri D, Sinks BC, Rauch SD, Shepard NT, White JA, et al. Utility of an abbreviated dizziness questionnaire to differentiate between causes of vertigo and guide appropriate referral: a multicenter prospective blinded study. *Otol Neurotol*. 2015, 36(10): 1687–1694.
8. Jacobson GP, Piker EG, Hatton K, Watford KE, Trone T, McCaslin DL, et al. Development and preliminary findings of the dizziness symptom profile. *Ear Hear*. 2019, 40(3): 568–578.
9. Friedland DR, Tarima S, Erbe C, Miles A. Development of a statistical model for the prediction of common vestibular diagnoses. *JAMA Otolaryngol Head Neck Surg*. 2016, 142(4): 351–356.

10. Yu F, Wu P, Deng H, Wu J, Sun S, Yu H, et al. A questionnaire-based ensemble learning model to predict the diagnosis of vertigo: model development and validation study. *J Med Internet Res*. 2022, 24(8): e34126.

11. Sushil M, Zack T, Mandair D, Zheng Z, Wali A, Yu YN, et al. A comparative study of large language model-based zero-shot inference and task-specific supervised classification of breast cancer pathology reports. *J Am Med Inform Assoc*. 2024, 31(10): 2315–2327.

12. Efron B, Tibshirani RJ. An introduction to the bootstrap. CRC Press; 1994.

13. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Springer; 2019.

14. Lampasona G, Piker E, Ryan C, Gerend P, Rauch SD, Goebel JA, et al. A systematic review of clinical vestibular symptom triage, tools, and algorithms. *Otolaryngol Head Neck Surg*. 2022, 167(1): 3–15.

15. Dennstädt F, Hastings J, Putora PM, Schmerder M, Cihoric N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *NPJ Digit Med*. 2025, 8(1): 143.