

Quantifying the Reasoning Abilities of LLMs on Clinical Cases (Supplementary Information)

Pengcheng Qiu^{1,2,*}, Chaoyi Wu^{1,2,*}, Shuyu Liu¹, Yanjie Fan³, Weike Zhao^{1,2}, Zhuoxia Chen⁴,
Hongfei Gu⁴, Chuanjin Peng⁴, Ya Zhang^{1,2}, Yanfeng Wang^{1,2,†} and Weidi Xie^{1,2,†}

¹Shanghai Jiao Tong University, Shanghai, China

²Shanghai Artificial Intelligence Laboratory, Shanghai, China

³Xin Hua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China

⁴China Mobile Communications Group Shanghai Co., Ltd., Shanghai, China

*Equal contributions †Corresponding author

Yanfeng Wang: wangyanfeng622@sjtu.edu.cn; Weidi Xie: weidi@sjtu.edu.cn

A Supplementary

A.1 Qualitative Case Study

In this section, the performance of various models is analyzed qualitatively to provide deeper insights, as shown in Supplementary Figure 1, 2, 3, 4, 5, 6, denoted as Case 1, 2, 3, 4, 5, 6, respectively. Additionally, we compare diagnostic and treatment evaluation standards to highlight differences in strictness levels, as shown in Supplementary Figure 7.

In cases, we will first show the ground-truth structured patient case, marked as “Patient Case”. It contains two identifiers:

- **Case ID:** The unique identifier assigned to each PMC case report in the dataset.
- **Category:** The classification framework for each case, encompassing affected body systems, specific disorders and conditions, and rare disease designations where applicable.

Then, to describe a certain patient case, it will further contain the following structured contents extracted and organized from the raw case reports through the model.

- **Case Summary:** Essential patient information including basic information and the results of ancillary tests. For Treatment Planning cases (Case 5), diagnostic conclusions are also included.
- **Reasoning Processes:** The documented clinical reasoning extracted from the original case report. In diagnosis cases (Case 2, 3, 4), this represents the differential diagnosis process. In treatment planning cases (Case 5), this delineates the therapeutic objectives and rationale underlying the selected intervention strategy.
- **Diagnosis or Treatment Results:** The definitive clinical outcomes as reported. For diagnostic cases (Case 2, 3, 4), this specifies the confirmed disease entity. For treatment cases (Case 5), this details the ultimately implemented therapeutic approach.

Afterward, the “Generation” part presents the outputs generated by different LLMs and the ground-truth for comparison, *i.e.*, examination recommendation (Case 1), diagnostic decision-making (Case 2, 3, 4), and treatment planning (Case 5). Notably, a special case arises in the context of examination recommendations, where the ground truth is not provided, as raw case reports rarely document the reasoning behind recommending specific examinations. Additionally, it is important to emphasize that Case 2 is continued after

Case 1, illustrating a scenario involving diagnostic decision-making after a 1-turn examination recommendation. Specifically, in Case 1, the models’ generated outputs conclude with the recommended examinations. Subsequently, in Case 2, a patient agent provides additional information in response, and the models are then tasked with making a diagnosis based on the newly provided information.

Lastly, we show the metrics for all cases, such as Accuracy, Precision, and Recall, and the definitions of them are consistent with the main body.

A.1.1 Examination Recommendation

Here, we analyze a case study in 1-turn examination recommendation, as shown in Case 1. Due to space constraints and to enhance clarity, we present the responses of only one open-source model (DeepSeek-R1) and one closed-source model (OpenAI-o3-mini).

In the initial phase of examination recommendation, encompassing the preliminary analysis and subsequent information inquiry, the models exhibited robust performance. They effectively identified pivotal symptoms, including vomiting, lethargy, skin darkening, and failure to thrive, utilizing these indicators to guide further information requests. The models demonstrated an understanding that specific symptom clusters were indicative of adrenal pathology, subsequently enumerating potential diagnoses such as congenital adrenal hyperplasia (CAH), Addison’s disease, and metabolic disorders. Building upon these tentative diagnoses, the models conducted a comprehensive assessment of the presenting symptoms to facilitate preliminary exclusion. For instance, DeepSeek-R1 ruled out neuroblastoma or other neoplastic conditions based on the absence of abdominal masses in the patient. In cases where definitive exclusion was not possible with the available information, the models considered requisite diagnostic tests for further examination. Regarding the presumptive diagnosis of CAH, the models contemplated relevant information that could refine diagnostic accuracy. Notably, they recognized the significance of normal genitalia as a critical piece of information, challenging the typical clinical presentation of CAH and prompting a more nuanced diagnostic approach.

Based on the former reasoning analysis, the models listed a comprehensive range of laboratory tests, including serum electrolytes, cortisol, adrenocorticotropic hormone (ACTH), 17-hydroxyprogesterone, renin, aldosterone levels, and blood glucose, imaging examinations (abdominal ultrasound), as well as genetic testing and urine steroid profile analysis. This underscores the model’s proficiency in identifying the requisite diagnostic-related examinations to differentiate among the various proposed diagnoses. Notably, during the evaluation phase, when contemplating Congenital Lipoid Adrenal Hyperplasia (CLAH), the model specifically solicited urine steroid profiles or genetic testing to ascertain the presence of adrenal enzyme deficiencies, thereby demonstrating a nuanced understanding of the specific disease and its features.

A.1.2 Diagnostic Decision-making

In this case study, we explore the diagnostic reasoning processes employed by DeepSeek-R1 and OpenAI-o3-mini across two different settings: (1) Diagnostic decision-making after 1-turn examination recommendation (Case 2). (2) Oracle diagnostic decision-making on a common disease case (Case 3) and a rare disease case (Case 4).

Diagnostic decision-making after 1-turn examination recommendation.

As shown in Case 2, when additional information response was presented by the patient agent, the assessed models carefully analyze each item. They noticed how the laboratory test results, such as hyponatremia, hyperkalemia, hypoglycemia, cortisol and aldosterone deficiencies, and low 17-hydroxyprogesterone, fit in with the previously considered diagnoses. It also took into account the genetic test result (homozygous mutation in the steroidogenic acute regulatory protein (StAR) gene) and the ultrasound finding (bilaterally enlarged adrenal glands) to further refine their understanding and inference. This shows that they could effectively integrate new information into the existing framework and adjust their thinking accordingly.

Then, based on the initial symptoms and available information, the model further excluded or retained the possibilities proposed in the initial analysis. For example, during the diagnostic reasoning process, OpenAI-o3-mini further excluded classic CAH according to the absence of acidosis in the patient, as metabolic acidosis is usually observed in classic CAH. DeepSeek-R1 also excluded 21-hydroxylase deficiency (classic CAH) based on the low 17-hydroxyprogesterone levels and the absence of virilization of the genitalia. Then, they used the

Patient Case		
Case ID	PMC11368709	
Category	- Body system: Endocrine System - Disorders and conditions: "Genetics/Birth Defects" - Common Disease-related	
Case Summary	Patient Information: - Patient Demographics: A three-week-old female neonate. - Chief Complaint: Vomiting, lethargy. - History of Present Illness: Progressive skin darkening and failure to thrive noted since birth. - Past Medical History: None. - Personal History: None. - Family History: No history of similar conditions, endocrine diseases, or neonatal deaths. - Physical Examination: Generalized dark skin pigmentation, normal female external genitalia, and no palpable abdominal or inguinal masses. Ancillary Tests: - Blood tests: Hypoglycemia (random blood sugar 28 mg/dL), persistent hyponatremia, hyperkalemia, and intermittent hypoglycemia despite therapy. - Ultrasound: Bilaterally enlarged adrenal glands and presence of Müllerian structures. - MRI of the brain: Occipital T2 hyperintensities indicating hypoglycemic insult. - Hormonal studies: Low 17-hydroxyprogesterone levels, cortisol deficiency, and aldosterone deficiency. - Genetic testing: Homozygous mutation in the steroidogenic acute regulatory protein (StAR) gene with autosomal recessive inheritance.	
Reasoning Processes	Differential diagnosis: 1. Initial considerations: - The patient's symptoms (vomiting, hyperpigmentation, hyponatremia, hyperkalemia, hypoglycemia, and failure to thrive) suggest adrenal insufficiency. - Differential diagnoses included common forms of congenital adrenal hyperplasia (CAH) such as 21-hydroxylase deficiency, 17,20 lyase deficiency, or 17-alpha-hydroxylase deficiency, as well as other causes of neonatal adrenal insufficiency like congenital lipid adrenal hyperplasia (CLAH). 2. Inclusion of CLAH: - Initial clinical clues emphasized generalized skin hyperpigmentation due to unchecked ACTH activity. - Persistently low levels of 17-hydroxyprogesterone differentiated CLAH from the more common 21-hydroxylase deficiency. - Enlarged adrenal glands on imaging were consistent with cholesterol accumulation seen in CLAH. - Genetic testing confirmed a homozygous StAR mutation, solidifying the diagnosis. 3. Exclusion of other causes: - Classic 21-hydroxylase deficiency was ruled out based on low 17-hydroxyprogesterone levels. - 17,20 lyase and 17-alpha-hydroxylase deficiencies were excluded due to the absence of other enzyme-specific deficiencies. - Sepsis and meningitis were ruled out early via initial investigations. 4. Final Confirmation: Hormonal studies, imaging, and genetic testing collectively confirmed congenital lipid adrenal hyperplasia due to a StAR mutation as the definitive diagnosis.	
Diagnosis Results	Congenital Lipoid Adrenal Hyperplasia (CLAH).	
Generation		
	DeepSeek-R1	OpenAI-o3-mini
Examination Recommendation	Reasoning: <step 1> The patient is a neonate with vomiting, lethargy, progressive skin darkening, and failure to thrive. Generalized hyperpigmentation suggests elevated ACTH levels, which may indicate adrenal insufficiency or congenital adrenal hyperplasia (CAH). However, normal female external genitalia in a genetic female makes classic salt-wasting CAH (21-hydroxylase deficiency) less likely, as this typically causes virilization of female genitalia. <step 2> Primary adrenal insufficiency (e.g., adrenal hypoplasia congenita, adrenal hemorrhage, or genetic defects in adrenal development) must be considered. These conditions can present with hyperpigmentation due to elevated ACTH, vomiting/lethargy from hyponatremia/hyperkalemia, and failure to thrive. The absence of family history does not exclude sporadic or X-linked recessive disorders. <step 3> Metabolic disorders (e.g., urea cycle defects, organic acidemias) could explain vomiting and lethargy, but hyperpigmentation is atypical. The lack of dysmorphic features or seizures reduces suspicion for many inborn errors of metabolism. <step 4> The absence of palpable abdominal/inguinal masses rules out adrenal tumors (e.g., neuroblastoma) or gonadal abnormalities. Normal genitalia also reduce suspicion for disorders of sexual development. <step 5> The constellation of hyperpigmentation, vomiting, and failure to thrive strongly points toward adrenal insufficiency. CAH with a less common enzyme deficiency (e.g., 3β-hydroxysteroid dehydrogenase deficiency, 11β-hydroxylase deficiency, or lipoid CAH) could still be possible despite normal genitalia, as these may present with milder or delayed virilization. Recommended Examinations: 1. Laboratory tests: Serum electrolytes (Na⁺, K⁺), glucose, cortisol, ACTH, 17-hydroxyprogesterone, renin, aldosterone. 2. Blood gas analysis to assess for metabolic acidosis. 3. Urine steroid profile or genetic testing for adrenal enzyme deficiencies. 4. Abdominal ultrasound to evaluate adrenal gland size/structure.	Reasoning: <step 1> Consider the patient demographics and chief complaint: The patient is a three-week-old female neonate presenting with vomiting, lethargy, and progressive skin darkening, along with failure to thrive since birth. These symptoms suggest a systemic condition rather than a local or isolated issue. <step 2> Evaluate the presenting symptoms in detail: - Vomiting and lethargy: Both of these symptoms are nonspecific but could indicate metabolic disturbances, dehydration, or a central nervous system issue. - Progressive skin darkening and failure to thrive since birth: Skin pigmentation changes from normal to dark could be indicative of several conditions, including metabolic or endocrine disorders like adrenal insufficiency, particularly congenital adrenal hyperplasia (CAH), or other disorders affecting pigmentation like hemochromatosis or certain mitochondrial diseases. <step 3> Assess the possible differential diagnoses: - Congenital Adrenal Hyperplasia (CAH): CAH, particularly 21-hydroxylase deficiency, is a leading concern due to its presentation with vomiting, lethargy, failure to thrive, and hyperpigmentation. The adrenal insufficiency caused by CAH leads to cortisol deficiency, which triggers an increase in ACTH, causing skin darkening (due to elevated melanocyte-stimulating hormone). It also typically presents in the neonatal period. - Hypothyroidism: Congenital hypothyroidism could lead to failure to thrive and lethargy, though skin darkening is not a classic feature. It is worth considering if metabolic tests reveal abnormalities. - Hemochromatosis: Although not common in neonates, certain metabolic disorders related to iron metabolism might cause skin pigmentation changes, but it would likely present with more specific signs like hepatomegaly or a history of iron overload. - Mitochondrial disorders: Certain mitochondrial diseases can present with failure to thrive, developmental delay, and skin pigmentation changes, but these typically have other neurologic manifestations and are rare. <step 4> Consider the family history and past medical history: The absence of any family history of endocrine diseases, similar conditions, or neonatal deaths may slightly reduce the likelihood of some inherited conditions (e.g., certain forms of hemochromatosis or metabolic disorders). However, congenital conditions like CAH can still occur sporadically without a family history. <step 5> Physical examination and signs: The generalized dark skin pigmentation is a key feature, particularly in the context of possible adrenal insufficiency, which is consistent with the diagnosis of CAH. The absence of palpable abdominal or inguinal masses does not suggest tumors or structural issues like neuroblastoma, which could present similarly. Recommended Examinations: 1. Laboratory tests: - Serum electrolytes: To assess for hyponatremia, hyperkalemia, and acidosis (common in CAH). - ACTH and cortisol levels: To confirm adrenal insufficiency. - 17-hydroxyprogesterone levels: This is the definitive test for diagnosing 21-hydroxylase deficiency, the most common form of CAH. - Renin and aldosterone levels: To evaluate for secondary adrenal insufficiency. 2. Imaging: - Ultrasound of the adrenal glands: While not always diagnostic, it can help assess for any abnormalities in the adrenal glands. 3. Genetic testing: If CAH is confirmed, genetic testing can help identify mutations in the CYP21A2 gene for definitive confirmation.
Metrics		
Precision	0.5	0.5
Recall	0.8	0.4

Supplementary Figure 1 | Case 1. A case of 1-turn examination recommendation. The meaning of different parts is explained at Supplementary Section [A.1](#)

Continue (PMC11368709)		
	DeepSeek-R1	OpenAI-o3-mini
Patient Agent Generation		
Patient-agent Provided Information	1. Laboratory Tests: - Serum electrolytes: Hyponatremia and hyperkalemia are noted. - Glucose: Persistent hypoglycemia reported, with a random blood sugar of 28 mg/dL. - Cortisol: Deficiency confirmed on hormonal studies. - ACTH: There is no relevant ancillary test information available for this request. - 17-hydroxyprogesterone: Low levels reported. - Renin: There is no relevant ancillary test information available for this request. - Aldosterone: Deficiency confirmed on hormonal studies. 2. Blood Gas Analysis: There is no relevant ancillary test information available for this request. 3. Urine Steroid Profile or Genetic Testing: Genetic Testing: Homozygous mutation in the steroidogenic acute regulatory protein (StAR) gene with autosomal recessive inheritance is reported. Urine Steroid Profile: There is no relevant ancillary test information available for this request. 4. Abdominal Ultrasound: Bilaterally enlarged adrenal glands are noted, with Müllerian structures present.	1. Laboratory Tests: - Serum electrolytes: The patient's blood tests show persistent hyponatremia and hyperkalemia. No mention of acidosis. - ACTH and cortisol levels: Hormonal studies confirm cortisol deficiency, though ACTH levels are not specifically mentioned. - 17-hydroxyprogesterone levels: Low 17-hydroxyprogesterone levels are reported, which is not typical for 21-hydroxylase deficiency (classic CAH), pointing toward an alternative diagnosis. - Renin and aldosterone levels: Hormonal studies indicate aldosterone deficiency. Renin levels are not reported. 2. Imaging: - Ultrasound of the adrenal glands: The ultrasound reveals bilaterally enlarged adrenal glands. 3. Genetic Testing: - Genetic testing shows a homozygous mutation in the steroidogenic acute regulatory protein (StAR) gene, with autosomal recessive inheritance. This finding is diagnostic for congenital lipid adrenal hyperplasia (CLAH), not 21-hydroxylase deficiency.
Generation		
Diagnostic Decision-Making	Reasoning: <step 1> The laboratory findings confirm adrenal insufficiency with hyponatremia, hyperkalemia, hypoglycemia, cortisol deficiency, and aldosterone deficiency. These findings align with a salt-wasting adrenal crisis. <step 2> Low 17-hydroxyprogesterone rules out 21-hydroxylase deficiency (classic CAH), which typically causes elevated 17-OHP and virilization of external genitalia in females. The absence of virilization here supports a non-classic CAH or other adrenal disorder. <step 3> Genetic testing reveals a homozygous mutation in the StAR gene, which is pathognomonic for lipoid congenital adrenal hyperplasia (lipoid CAH). StAR deficiency impairs cholesterol transport into mitochondria, disrupting steroidogenesis in both adrenals and gonads. <step 4> Bilaterally enlarged adrenal glands on ultrasound are consistent with lipoid CAH, as cholesterol accumulation leads to adrenal hyperplasia. Normal female genitalia and Müllerian structures align with StAR deficiency, as impaired gonadal steroidogenesis prevents fetal androgen production, sparing external genitalia in genetic females. <step 5> The combination of adrenal insufficiency, low 17-OHP, StAR mutation, and adrenal hyperplasia confirms lipoid CAH. Other adrenal disorders (e.g., adrenal hypoplasia congenita) are excluded due to adrenal enlargement and specific genetic findings. Disease Diagnosis: Lipoid Congenital Adrenal Hyperplasia (StAR Deficiency)	Reasoning: <step 1> Review of the initial symptoms and findings: The patient, a three-week-old neonate, presented with vomiting, lethargy, failure to thrive, and progressive skin darkening since birth. These symptoms, combined with laboratory findings and imaging results, strongly point toward an adrenal disorder. <step 2> Re-examine laboratory findings: - The presence of persistent hyponatremia and hyperkalemia in the patient is indicative of adrenal insufficiency, which is a common feature in disorders like congenital adrenal hyperplasia (CAH). However, the absence of acidosis does not strongly support classic CAH, where metabolic acidosis is typically observed. - Cortisol deficiency was confirmed, aligning with adrenal insufficiency. The low 17-hydroxyprogesterone levels are atypical for 21-hydroxylase deficiency, suggesting the absence of classic CAH. This directs us toward other forms of adrenal insufficiency. - Aldosterone deficiency further supports the diagnosis of an adrenal disorder, as aldosterone is a key hormone produced by the adrenal glands. The absence of mention of renin levels, while helpful, does not change the overall impression. <step 3> Ultrasound results of the adrenal glands: The bilateral enlargement of the adrenal glands is a significant clue. In conditions like congenital adrenal hyperplasia, adrenal glands may be enlarged due to excessive ACTH stimulation. In this case, the enlargement is also compatible with congenital lipoid adrenal hyperplasia (CLAH), which is associated with steroidogenic enzyme deficiencies and adrenal enlargement. <step 4> Genetic findings: The genetic testing revealed a homozygous mutation in the steroidogenic acute regulatory protein (StAR) gene, which is diagnostic of congenital lipoid adrenal hyperplasia (CLAH). This is a rare but severe form of adrenal insufficiency where the defective StAR protein impairs cholesterol transport into the mitochondria, leading to an inability to produce steroid hormones like cortisol and aldosterone. Disease Diagnosis: congenital lipoid adrenal hyperplasia (CLAH)
Metrics		
Accuracy	1	1

Supplementary Figure 2 | Case 2. A case of 1-turn diagnostic decision-making. It is a continue figure with Supplementary Figure 1. The meaning of different parts is explained at Supplementary Section A.1

genetic test result and the ultrasound finding to confirm the diagnosis of lipoid congenital adrenal hyperplasia (CLAH). This indicates that the models could use information in a logical and evidence-based manner to reach the correct final diagnosis.

Oracle diagnostic decision-making.

When provided with complete information, LLMs tend to perform accurate diagnosis. In this context, we will analyze deeper into evaluating the efficiency of reasoning, the factuality of each step, the completeness (recall of ground-truth reasoning steps), and the final diagnostic accuracy.

Efficiency: DeepSeek-R1’s approach is characterized by its conciseness and directness. In Case 3, it swiftly progresses from symptoms to targeting Traboulsi syndrome through genetic testing and clinical manifestations. Similarly, in Case 4, it efficiently eliminates malignant and infectious causes early on, focusing on mass characteristics and ultrasound findings for diagnosis. In contrast, OpenAI-o3-mini provided a more detailed step-by-step analysis. In Case 3, it involved more physical examinations and eye findings. In Case 4, it emphasized the benign nature of the mass through the absence of symptoms typically associated with malignancy or infection. While this information contributes to a comprehensive understanding of the condition, it also adds complexity to reasoning.

Factuality: While generally aligned with medical knowledge, there are instances where the models deviate. In Case 3, DeepSeek-R1’s reasoning step 4 is not strictly factual. The absence of cardiovascular or metabolic abnormalities is not a sufficient condition to exclude Marfan or Weill-Marchesani syndromes, which are characterized by specific symptoms like aortic enlargement or joint hypermobility, and short stature or brachydactyly, respectively. Meanwhile, the assertion in step 5 of OpenAI-o3-mini’s reasoning regarding mutations in the ASPH gene lacks medical support and may have significantly contributed to its incorrect final diagnosis. In Case 4, however, both models adhere closely to medical principles, with DeepSeek-R1 correctly associating symptom absence with a benign process and OpenAI-o3-mini accurately emphasizing the importance of mass characteristics in suggesting a benign etiology.

Completeness: DeepSeek-R1 demonstrates a strong ability to recall and follow ground-truth reasoning steps in both cases, particularly in identifying the key clues such as the benign nature of lesions and the importance of gene testing or histopathology in diagnosis. OpenAI-o3-mini, however, shows a limited capacity to exclude differential diagnoses explicitly. In Case 3, it failed to clearly rule out Weill-Marchesani syndrome, and in Case 4, it did not definitively exclude Bartholin cyst, despite its detailed step-by-step analysis.

Diagnostic Accuracy: In terms of final diagnosis, both models correctly identify the ASPH mutation in Case 3. However, OpenAI-o3-mini misdiagnoses the condition as "Asperger Syndrome (ASPH gene mutation-related disorder)", which is a misinterpretation, as Asperger Syndrome is a pervasive developmental disorder not proven to be related to the ASPH gene. This indicates a gap in OpenAI-o3-mini’s medical knowledge and reasoning rigor. In Case 4, both DeepSeek-R1 and OpenAI-o3-mini successfully diagnosed vulvar leiomyoma. This consistency with the ground-truth suggests that both models are capable of accurate clinical reasoning when presented with clear symptoms and diagnostic findings.

A.1.3 Treatment Planning

Here, we present a case of a rare disease and analyze the treatment planning processes employed by Deepseek-R1, illustrated in Case 5. We focus on analyzing the differences between the reasoning generated in the formal answers (for simplicity, denoted as Chain-of-thought) and the default thinking process generated by Deepseek-R1. Our analysis is grounded in evaluating the efficiency of reasoning, the factuality of each step, and the completeness.

First, we provide a brief overview of the case. The patient is a 68-year-old male diagnosed with cutis laxa Ehlers-Danlos syndrome (dEDS). His main symptoms include extremely fragile skin, hypermobile joints, and high blood pressure. He has a known allergy to sulfonamides, which previously caused a systemic rash. Regarding his medication status, he was prescribed Triamterene-hydrochlorothiazide (37.5 mg/25 mg daily) to manage his hypertension. Additionally, he has a family history of chronic kidney disease.

Efficiency: We investigated the effectiveness in the chain-of-thought and the thinking process reasoning. Overall, the output of chain-of-thought is more concise, while that of the thinking process is more fragmented.

Patient Case			
Case ID	PMC11431244		
Category	- Body system: Eyes and Vision - Disorders and conditions: "Genetics/Birth Defects" - Common Disease-related		
Case Summary	Patient Information: - Patient Demographics: 21-year-old Mexican male. - Chief Complaint: Trauma to the right eye while playing soccer. - History of Present Illness: Blurred vision since childhood, previously diagnosed with lens subluxation and dysmorphic facial features, but no genetic testing conducted. - Past Medical History: None mentioned. - Personal History: Born via cesarean section at 41 weeks; no notable medical history; no consanguinity. - Family History: No significant genetic or ophthalmic disorders reported. - Physical Examination: Dysmorphic facial features (elongated face, downward slanting palpebral fissures, prominent nose, nasal alae hypoplasia, high-arched palate, and crowded teeth); scoliosis present. Ancillary Tests: - Ophthalmic evaluation: Right eye (OD) lens subluxation into the anterior chamber with iris atrophy and glaukomflecken; left eye (OS) posterior and inferior lens subluxation with sectorial iris atrophy and iridodonesis; both eyes exhibited small, pale optic discs with a tessellated fundus. - Genetic tests: Compound heterozygous ASPH gene variants identified. One known pathogenic variant (NM_004318.3:c.1892G>A, p.Trp631*) and one novel likely pathogenic variant (deletion of exons 20–21).		
Reasoning Processes	Differential diagnosis: 1. Initial considerations: - The patient's symptoms of lens subluxation, facial dysmorphism, and visual disturbances raised suspicion for connective tissue disorders such as Marfan syndrome and other ectopia lentis syndromes (e.g., Weill–Marchesani syndrome). 2. Why Marfan syndrome was excluded: - The absence of systemic features such as aortic enlargement or joint hypermobility and the presence of dysmorphic facial features specific to Traboulsi syndrome made Marfan syndrome unlikely. 3. Why Weill–Marchesani syndrome was excluded: - Unlike Weill–Marchesani syndrome, which exhibits short stature and brachydactyly, this patient had features more consistent with Traboulsi syndrome (e.g., facial dysmorphism and anterior segment anomalies). 4. Confirmation of Traboulsi syndrome: - Genetic testing confirmed pathogenic ASPH gene variants, hallmark signs of Traboulsi syndrome. The distinct ocular findings, including lens subluxation and sectorial iris atrophy without postoperative filtering blebs, were consistent with phenotypic variability associated with Traboulsi syndrome.		
Diagnosis Results	Traboulsi syndrome		
Generation			
	Ground Truth	DeepSeek-R1	OpenAI-o3-mini
Diagnostic Decision-Making	Reasoning: <Step 1> The patient's symptoms of lens subluxation, facial dysmorphism, and visual disturbances raised suspicion for connective tissue disorders such as Marfan syndrome and other ectopia lentis syndromes (e.g., Weill–Marchesani syndrome). <Step 2> Marfan syndrome was considered unlikely due to the absence of systemic features such as aortic enlargement or joint hypermobility and the presence of dysmorphic facial features specific to Traboulsi syndrome. <Step 3> Weill–Marchesani syndrome was excluded because the patient did not exhibit short stature or brachydactyly, which are key features of this syndrome, and instead had features more consistent with Traboulsi syndrome. <Step 4> Genetic testing confirmed pathogenic ASPH gene variants, which are hallmark signs of Traboulsi syndrome. <Step 5> The distinct ocular findings, including lens subluxation and sectorial iris atrophy without postoperative filtering blebs, were consistent with the phenotypic variability associated with Traboulsi syndrome. <Step 6> Traboulsi syndrome was confirmed through the identification of compound heterozygous pathogenic ASPH variants and supported by the patient's unique clinical presentation, including lens subluxation, ocular abnormalities, dysmorphic facial features, and phenotypic variability. Disease Diagnosis: Traboulsi syndrome	Reasoning: <step 1> The patient has a history of lens subluxation and dysmorphic facial features (elongated face, downward slanting palpebral fissures, nasal alae hypoplasia, high-arched palate, crowded teeth), along with scoliosis. These features suggest a systemic connective tissue disorder with ocular involvement. <step 2> Genetic testing revealed compound heterozygous ASPH gene variants. ASPH mutations are strongly associated with Traboulsi syndrome (ectopia lentis syndrome 2, OMIM #618615), an autosomal recessive disorder characterized by ectopia lentis (lens subluxation), distinct craniofacial dysmorphism, and skeletal abnormalities like scoliosis. <step 3> The pathogenic variant (c.1892G>A, p.Trp631*) is a nonsense mutation leading to premature termination, while the deletion of exons 20–21 likely disrupts ASPH's calcium-binding epidermal growth factor-like domains, critical for protein function. Compound heterozygosity confirms the autosomal recessive inheritance pattern. <step 4> The ophthalmic findings (lens subluxation, iris atrophy, glaukomflecken, small optic discs) align with ASPH-related ocular pathology. The absence of cardiovascular or metabolic abnormalities rules out other ectopia lentis-associated syndromes (e.g., Marfan, Weill–Marchesani) Disease Diagnosis: Traboulsi syndrome (ASPHD-related ectopia lentis syndrome) due to compound heterozygous pathogenic ASPH variants.	Reasoning: <step 1> The patient is a 21-year-old male with a history of blurred vision since childhood. This suggests a long-standing, possibly progressive ophthalmic condition, which needs to be considered in the differential diagnosis. The patient's previous diagnosis of lens subluxation is notable, as it is a key finding associated with certain genetic conditions. <step 2> The presence of dysmorphic facial features, including an elongated face, downward slanting palpebral fissures, prominent nose, nasal alae hypoplasia, high-arched palate, and crowded teeth, points towards a genetic syndrome with a connective tissue or skeletal dysplasia component. These features are commonly seen in syndromes such as Marfan syndrome, Homocystinuria, and others, which involve connective tissue abnormalities and ocular manifestations. <step 3> The physical findings also include scoliosis, which is a musculoskeletal manifestation that often correlates with systemic conditions affecting the connective tissue, further suggesting a potential syndromic diagnosis. <step 4> Ophthalmic findings are crucial: - The right eye shows lens subluxation into the anterior chamber with iris atrophy and glaukomflecken. - The left eye shows posterior and inferior lens subluxation with sectorial iris atrophy and iridodonesis. - Both eyes have small, pale optic discs with a tessellated fundus. These features of lens subluxation, iris atrophy, and glaukomflecken are highly suggestive of a connective tissue disorder with ocular involvement, such as homocystinuria or Marfan syndrome. However, the specific pattern of these findings, including both iris atrophy and glaukomflecken, may be more indicative of a different condition like a genetic connective tissue disorder linked to ASPH gene mutations. <step 5> Genetic testing revealed compound heterozygous ASPH gene variants, one known pathogenic (NM_004318.3:c.1892G>A, p.Trp631*) and one novel likely pathogenic variant (deletion of exons 20–21). ASPH gene mutations are associated with a rare genetic disorder called "aspartylglucosaminuria," which often involves dysmorphic features, lens subluxation, and other systemic manifestations like scoliosis. Disease Diagnosis: Asperger syndrome (ASPH gene mutation-related disorder)
Metrics			
Accuracy	\	1	0
Efficiency	\	1	1
Factuality	\	0.75	0.80
Recall	\	1	0.83

Supplementary Figure 3 | Case 3. A case of oracle diagnosis on common disease. The meaning of the row headers is explained at the beginning of Supplementary A.1

Patient Case			
Case ID	PMC11407790		
Category	- Body system: Others - Disorders and conditions: “ Cancers” - Rare Disease-related		
Case Summary	Patient Information: - Patient Demographics: 40-year-old woman - Chief Complaint: Painless, gradually enlarging mass on the left side of the vulva - History of Present Illness: Swelling progressively increasing in size, slight discomfort, no associated symptoms such as discharge, bleeding, or urinary changes. - Past Medical History: None provided - Personal History: None provided - Family History: None provided - Physical Examination: A solid, well-circumscribed, non-tender lump (3 x 2 cm) on the left vulva. Ancillary Tests: - Complete blood count: Hemoglobin 11.2 g/dl, white blood cell count 4,500 cells/cu.mm, platelet count 4,72,000 cells/cu.mm - Pelvic ultrasonography: Firm, hypochoic mass consistent with a benign lesion. - Excisional biopsy: Single, whitish, firm mass (3.5 x 3 x 1.8 cm) analyzed histopathologically. Diagnosis confirmed as vulval leiomyoma based on interlacing smooth muscle bundles without atypia or mitotic activity.		
Reasoning Processes	Differential diagnosis: 1. Initial Considerations: - Based on the patient's presentation of a painless, slow-growing vulvar mass, initial considerations included benign vulval lesions such as Bartholin cyst, lipoma, or fibroma. Malignant causes were less likely based on the absence of acute symptoms, ulceration, or systemic signs. 2. Ruling Out Diagnoses: - Bartholin Cyst: Typically associated with a fluid-filled cystic appearance and located more inferiorly, which did not match the findings of a firm, hypochoic solid mass in this case. - Lipoma: Generally presents as a softer, fatty lump, inconsistent with the hypochoic characteristic seen on ultrasound. - Other possibilities (e.g., fibromas): Could not be ruled out until histopathological confirmation. 3. Confirmation via Biopsy: - The excisional biopsy was definitive, showing characteristic histopathological features of vulval leiomyoma (interlacing bundles of smooth muscle cells with no atypia or mitotic activity).		
Diagnosis Results	Vulval Leiomyoma		
Generation			
	Ground Truth	DeepSeek-R1	OpenAI-o3-mini
Diagnostic Decision-Making	Reasoning: <Step 1> The patient's presentation of a painless, slow-growing vulvar mass initially suggested benign vulval lesions such as Bartholin cyst, lipoma, or fibroma, with malignant causes being less likely due to the absence of acute symptoms, ulceration, or systemic signs. <Step 2> Bartholin cyst was ruled out because it is typically associated with a fluid-filled cystic appearance and is located more inferiorly, which did not match the findings of a firm, hypochoic solid mass seen in this case. <Step 3> Lipoma was ruled out because it generally presents as a softer, fatty lump, inconsistent with the hypochoic characteristic seen on ultrasound. <Step 4> Fibroma and other possibilities could not be definitively ruled out without histopathological confirmation. <Step 5> An excisional biopsy was performed, which revealed characteristic histopathological features of vulval leiomyoma, including interlacing bundles of smooth muscle cells with no atypia or mitotic activity. <Step 6> The patient was diagnosed with vulval leiomyoma, a rare benign smooth muscle tumor of the vulva, based on the histopathological examination. Disease Diagnosis: Vulval Leiomyoma	Reasoning: <step 1> The patient presents with a painless, slowly enlarging vulvar mass. The absence of symptoms like bleeding, discharge, or urinary changes suggests a benign process rather than infection or malignancy. <step 2> Physical examination reveals a solid, well-circumscribed, non-tender mass. These features are consistent with benign tumors (e.g., leiomyoma, fibroma) rather than malignant lesions, which often exhibit irregular borders, fixation, or ulceration. <step 3> Pelvic ultrasonography identifies a hypochoic mass, aligning with a solid benign lesion. This supports excluding cystic pathologies (e.g., Bartholin’s cyst) or vascular abnormalities. <step 4> Histopathological analysis of the excised mass confirms interlacing smooth muscle bundles without atypia or mitotic activity. This definitively rules out malignancy (e.g., leiomyosarcoma) and establishes the diagnosis of a smooth muscle neoplasm. <step 5> The combination of clinical presentation, imaging findings, and histopathology confirms vulval leiomyoma, a rare benign tumor arising from smooth muscle cells in the vulva. Disease Diagnosis: Vulval Leiomyoma	Reasoning: <step 1> The patient's chief complaint is a painless, gradually enlarging mass on the left side of the vulva. The key characteristics here are that the mass is painless, non-tender, and gradually growing in size. This suggests a benign process as malignant growths tend to be more rapid in their progression and often cause discomfort. <step 2> The absence of symptoms like discharge, bleeding, or urinary changes further supports a benign etiology. If the mass were malignant or infectious, one would expect some associated systemic or local symptoms. <step 3> The physical examination describes a solid, well-circumscribed, non-tender lump (3 x 2 cm) on the left vulva. The well-circumscribed nature and non-tenderness of the lump are also indicative of a benign mass, as malignant tumors tend to be irregular in shape and tender. <step 4> The pelvic ultrasonography findings of a firm, hypochoic mass that is consistent with a benign lesion suggest a non-cystic mass, which supports the possibility of a benign tumor like a fibroid or leiomyoma. <step 5> The excisional biopsy result reveals a single, whitish, firm mass, with histopathology confirming the diagnosis of vulval leiomyoma. Leiomyomas are benign smooth muscle tumors, often found in the vulva, and this finding is consistent with the clinical presentation. Disease Diagnosis: Vulval Leiomyoma
Metrics			
Accuracy	\	1	1
Efficiency	\	1	0.80
Factuality	\	1	1
Recall	\	1	0.67

Supplementary Figure 4 | Case 4. A case of oracle diagnosis on rare diseases. The meaning of different parts is explained at Supplementary Section [A.1](#)

Patient Case			
Case ID	PMC11624969		
Category	- Body system: "Blood, Heart and Circulation", "Bones, Joints and Muscles", "Skin, Hair and Nails" - Disorders and conditions: "Genetics/Birth Defects" - Rare Disease-related.		
Case Summary	- Patient Demographics: A 68-year-old Caucasian male - Chief Complaint: Establishing care; no acute distress reported. - History of Present Illness: - Diagnosed with dermatosparaxis-type Ehlers-Danlos Syndrome (dEDS), characterized by extreme skin fragility, redundant lax skin, multiple bruises (petechiae and hematomas), puffy eyelids, epicanthal folds, and generalized joint hypermobility. - Currently has controlled hypertension, primary osteoarthritis of the right knee, and history of carpal tunnel syndrome in the right wrist. - Medications: Started triamterene-hydrochlorothiazide (37.5-25 mg daily) for hypertension, resulting in well-controlled blood pressure. - Past Medical History: - Primary hypertension, primary right knee osteoarthritis, carpal tunnel syndrome (right wrist). - No atrophic scars or poor wound healing despite significant skin fragility. - Allergies: Sulfa drugs, previously causing a generalized rash. - Family History: Father - cancer, chronic kidney disease; mother - cancer. - Social History: Former cigarette smoker, quit 11 years ago. - Physical Exam: - Blood pressure: 142/89 mmHg at presentation (controlled with medication later), heart rate: 87 bpm, BMI: 30.12. - Generalized joint hypermobility; dermatologic abnormalities include redundant, hyperextensible, soft, and doughy skin. Evidence of petechiae and subcutaneous hematomas. - Lab Events: - Normal complete blood count, basic metabolic panel, and lipid panel. - Ambulatory blood pressure monitoring showed consistently elevated diastolic blood pressure before treatment. - Imaging Events: - Duplex ultrasound negative for abdominal aortic aneurysm or aortic dilation. - 12-lead ECG negative for left atrial enlargement, left ventricular hypertrophy, or mitral valve prolapse. - Final Diagnostic Results: Clinical diagnosis of dEDS confirmed based on extreme skin fragility, craniofacial features, joint hypermobility.		
Reasoning Processes	Treatment goals and rationale for the chosen interventions: Objective: Firstly, address dermatologic, orthopedic, and cardiovascular complications of dermatosparaxis-type Ehlers-Danlos Syndrome. Secondly, focus on improving physical function, preventing complications (e.g., joint injuries, bruising), and maintaining overall cardiovascular health. Rationale: (1) Skin fragility management is crucial to prevent traumatic injuries and enhance healing. Recommendations include protective measures and avoidance of high-risk activities. (2) Orthopedic manifestations such as joint hypermobility and osteoarthritis require tailored physical therapy and possibly supportive devices to avoid instability and long-term joint complications. (3) Cardiovascular risks related to vessel fragility and potential for aortic dilation necessitate regular imaging surveillance and blood pressure control. (4) Multidisciplinary care is required for specialized support in symptom management and overall quality of life improvement for this rare connective tissue disorder.		
Treatment Results	Enhance management by utilizing protective measures, engaging in low-impact physical therapy, monitoring cardiovascular health regularly, and administering antihypertensive therapy as prescribed.		
Generation			
	Ground Truth	DeepSeek-R1 (Chain of thought)	DeepSeek-R1 (Thinking process)
Treatment Planning	Reasoning: <Step 1> The objective is to address dermatologic, orthopedic, and cardiovascular complications of dermatosparaxis-type Ehlers-Danlos Syndrome. <Step 2> The treatment plan aims to improve physical function, prevent complications such as joint injuries and bruising, and maintain overall cardiovascular health. <Step 3> Skin fragility is managed through protective measures and avoidance of high-risk activities to prevent traumatic injuries and enhance healing. <Step 4> Orthopedic complications, including joint hypermobility and osteoarthritis, are addressed with tailored physical therapy and supportive devices to avoid joint instability and long-term complications. <Step 5> Cardiovascular risks stemming from vessel fragility and aortic dilation are managed with regular imaging surveillance and blood pressure control. <Step 6> Multidisciplinary care is emphasized to provide specialized support, manage symptoms effectively, and improve the overall quality of life for patients with this rare connective tissue disorder. <Step 7> The treatment plan incorporates protective measures, low-impact physical therapy, regular cardiovascular health monitoring, and antihypertensive therapy as prescribed. Treatment Plans: (Refer to Treatment Results)	Reasoning: <Step 1> Assess the patient's current medications for safety and compatibility with allergies. The patient has a sulfa allergy (generalized rash), and hydrochlorothiazide (HCTZ) is a sulfonamide derivative. Despite tolerating it currently, there is a risk of cross-reactivity. Switching to a non-sulfa antihypertensive is prudent to avoid potential allergic complications. <Step 2> Evaluate hypertension management alternatives. ACE inhibitors (e.g., lisinopril) or ARBs (e.g., losartan) are appropriate replacements. These agents also offer renal protective benefits, aligning with the patient's family history of chronic kidney disease. <Step 3> Address dermatosparaxis-specific care: prioritize skin protection (e.g., padded clothing, avoiding trauma), dermatology referrals for proactive monitoring, and gentle wound care protocols to mitigate complications from skin fragility. <Step 4> Manage osteoarthritis and joint hypermobility: Avoid NSAIDs (risk of bruising/gastrointestinal bleeding). Recommend acetaminophen for pain, physical therapy for joint stabilization, and weight reduction to alleviate knee stress. <Step 5> Monitor for carpal tunnel recurrence and cardiovascular risks. Ensure regular follow-ups for blood pressure control, renal function, and skin/joint health. Treatment Plans: Discontinue hydrochlorothiazide-triamterene; initiate lisinopril for hypertension. Implement skin protection measures, physical therapy for joint stabilization, and acetaminophen for osteoarthritis pain. Regular dermatology and primary care follow-ups.	Reasoning: (1) Okay, let's start by understanding the patient's case. He's a 68-year-old man with dermatosparaxis-type Ehlers-Danlos Syndrome (dEDS). His main issues are skin fragility, joint hypermobility, and controlled hypertension. He's on triamterene-HCTZ for blood pressure, which is working well. But he's also got osteoarthritis in his right knee and a history of carpal tunnel. (2) First, I need to focus on managing his dEDS. Since his skin is super fragile, preventing injuries and skin tears is crucial. Maybe recommend protective clothing or padding. Also, physical therapy for joint stability could help with hypermobility and osteoarthritis. But have to be careful with exercises to avoid dislocations. (3) Looking at his meds, he's on hydrochlorothiazide. Wait, he has a sulfa allergy. Hydrochlorothiazide is a sulfonamide, right? Oh, but he's been taking it without issues. The allergy was a rash from sulfa drugs before. Hmm, but sometimes there's cross-reactivity. Maybe consider switching to a different antihypertensive to avoid potential allergic reactions. Like an ACE inhibitor or ARB, which are also good for kidney protection, especially since his family has CKD history. (4) His blood pressure was 142/89 initially but controlled now. Need to keep monitoring that. ACE inhibitors can help with both hypertension and possibly kidney function. But check his potassium levels because triamterene is a potassium-sparing diuretic. If switching from HCTZ, might need to adjust other meds. (5) For osteoarthritis, usual treatments like acetaminophen, but NSAIDs might be risky due to his skin fragility and bruising. Maybe avoid NSAIDs to prevent gastrointestinal issues or bleeding. Physical therapy and weight management could help his knee. His BMI is 30.12, so weight loss might reduce joint stress. (6) Carpal tunnel history—maybe assess if he needs splinting or if symptoms recur. But since it's past, just monitor. (7) Follow-up for skin issues: regular dermatology check-ups. Educate on wound care, even though he hasn't had poor healing yet. Genetic counseling might be beneficial for family planning aspects, but at 68, maybe less relevant. (8) Labs are normal, so no immediate issues there. Continue monitoring kidney function, especially with hypertension meds and family history. (9) So, key points: switch hypertension medication away from sulfa-containing HCTZ due to allergy, manage dEDS symptoms preventatively, address osteoarthritis without NSAIDs, and monitor overall health.
Metrics			
Accuracy	\	1	1
Efficiency	\	1	0.89
Factuality	\	1	1
Recall	\	0.43	1

Supplementary Figure 5 | Case 5. A case of treatment planning for a rare disease. The meaning of different parts is explained at Supplementary Section A.1.

From the cases, we can see that the thinking process divides the reasoning of the entire problem into 9 steps, whereas Deepseek-R1 only used 5 steps. Additionally, in the first step, the thinking process initially summarized the entire medical record without beginning the reasoning. This is considered an ineffective reasoning step. Therefore, the efficiency of the thinking process is slightly lower, at only 0.89, while the efficiency of chain-of-thought is 1.

Factuality: In this case, the chain-of-thought and thinking process are largely true and logically sound, with each step building on relevant clinical considerations. The recommendations for managing dEDS, including prioritizing skin protection, dermatology referrals, and gentle wound care protocols, are appropriate given the patient’s skin fragility. The approach to osteoarthritis management, which avoids NSAIDs due to the risk of bruising and gastrointestinal bleeding, is also reasonable, especially considering the patient’s specific condition. The suggestion to switch hypertension medication away from HCTZ due to the patient’s sulfa allergy is cautious and aligns with the potential risk of cross-reactivity, even though the patient has tolerated it thus far. While some plans may be overly cautious or lack specificity, such as the broad suggestion to monitor overall health, the overall reasoning is factually correct and follows a logical progression.

Completeness: We find that Deepseek-R1’s thinking process is more complete than the chain-of-thought. The thinking process covers most of the ground-truth rationale, while the chain-of-thought misses some key thinking procedures during its process. Specifically, the chain-of-thought does not explicitly address the overarching objectives of managing dermatosparaxis-type Ehlers-Danlos Syndrome (dEDS), including the comprehensive management of dermatologic, orthopedic, and cardiovascular complications (<Step 1>). Additionally, it fails to outline the treatment plan’s goals of improving physical function, preventing joint injuries and bruising, and maintaining cardiovascular health (<Step 2>). The chain-of-thought also neglects the management of cardiovascular risks associated with vessel fragility and aortic dilation, such as regular imaging surveillance and blood pressure control (<Step 5>), and does not emphasize the importance of multidisciplinary care for specialized support, effective symptom management, and overall quality-of-life improvement (<Step 6>). This leads to a relatively low completeness score.

Treatment Planning Accuracy: In terms of the final treatment planning, Deepseek-R1 provided a rational treatment plan. In this case, Deepseek-R1’s output was largely consistent with the ground truth. However, it introduced two novel treatment recommendations: first, to discontinue Triamterene-hydrochlorothiazide and initiate Lisinopril for hypertension management; second, to use Paracetamol for alleviating osteoarthritis pain. Our evaluation pipeline, which integrates a search engine to gather relevant information, confirmed that these recommendations are reasonable.

Considering the patient’s family history of chronic kidney disease, switching to Lisinopril is a prudent choice, as it aligns with clinical guidelines for managing hypertension in such contexts. Additionally, given the patient’s known allergy to sulfonamides, discontinuing Triamterene-hydrochlorothiazide is justified due to the potential for cross-allergic reactions. Furthermore, the recommendation to use Paracetamol for osteoarthritis pain is appropriate, as it is a commonly accepted treatment option for this condition.

Therefore, based on these evaluation findings, the final judgment is that the treatment plan proposed by Deepseek-R1 is correct and clinically sound.

A.1.4 Rare Disease Cases with Specific Tests

We also dive into the reason why model’s performance on diagnostic decision making on rare diseases doesn’t decrease significantly compared to that on common diseases, considering that rare diseases is a much more harder task for most of physicians. We carefully check our dataset and find that which is a typical example where a specific test can lead to final diagnostic decision. In oracle settings, where the LLM is provided with all the necessary information, including auxiliary examinations, the task becomes relatively straightforward. Meanwhile, in dynamic diagnostic decision-making, where the LLM is required to actively query the appropriate auxiliary examinations, the task becomes more challenging, as some LLMs fail to identify or request the specific examinations needed. However, once query the adequate examination such as the specific task, the diagnose become easy.

Here, we present a case of diagnostic decision-making for a rare disease, as illustrated in Figure 6. In this case, the patient was diagnosed with adult nephroblastoma, a rare disease that accounts for less than 1% of all renal

malignancies in adults. The patient’s specific diagnostic test is the histopathological confirmation of a triphasic pattern (blastemal, stromal, and epithelial components), which directly leads to the final diagnostic decision. During the diagnostic reasoning process, the LLMs also identified this key clue and quickly arrived at the correct final decision. For DeepSeek-V3, it recognized that the histopathological confirmation of a triphasic pattern (blastemal, stromal, and epithelial components) is diagnostic of nephroblastoma (Wilms tumor) in Step 2. Similarly, OpenAI-o3-mini identified that the histopathology report confirmed a nephroblastoma (Wilms tumor) with a triphasic pattern involving blastemal, stromal, and epithelial components in Step 3. Ultimately, both models successfully provided the correct diagnostic result.

A.1.5 Comparison of diagnostic and treatment evaluation standards

As shown in Supplementary Figure 7, a diagnosis is considered correct if it accurately identifies the primary disease and metastatic pattern. In contrast, the treatment plan is deemed incorrect—even if it includes embolization—because deviations in target vessel selection or procedural technique fail to meet the stricter evaluation criteria for treatment planning.

This demonstrates that treatment accuracy is held to higher standards than diagnostic accuracy, reflecting the greater complexity of treatment planning.

A.2 Generation Parameters

For all evaluated models in this study, the default temperature was set to 1.0, except for DeepSeek-R1, where a temperature of 0.6 was used based on its official documentation, which states that 0.6 yields the best performance. When utilizing LLM as judge in the evaluation system, the temperature was uniformly set to 0.0 to minimize randomness. Additionally, the random seed was fixed at 42 to ensure reproducibility.

Patient Case			
Case ID	PMC11492433		
Category	- Body system: Kidneys and Urinary System - Disorders and conditions: Cancers - Rare Disease-related		
Case Summary	Patient Information: - 48-year-old male - Chief Complaint: Persistent left abdominal pain - History of Present Illness: Six-month history of progressively worsening left abdominal pain, with associated unintentional weight loss and increasing fatigue. - Past Medical History: None noted (non-contributory) - Personal History: None provided - Family History: None mentioned - Physical Examination: No specific physical abnormalities beyond the chief complaint of abdominal pain. Ancillary Tests: - Pre-operative laboratory tests: Mild anemia (Hb 10.5 g/dL) without other hematologic or metabolic abnormalities. Renal and liver function tests were normal. - Imaging: CT scan revealed a large (12 cm), heterogeneous mass in the left kidney with solid and cystic components, irregular borders, enlarged regional peri-hilar and left aortic lymph nodes (1.2 cm), and suspected local extension into perinephric fat. No distant metastases identified. - Histopathology: Post-nephrectomy histology confirmed nephroblastoma with a triphasic pattern (blastemal, stromal, and epithelial components). Regional lymph nodes exhibited metastatic involvement. Pathologic staging: pT2b N2 (local extension to the peri-hilar region with metastatic node involvement). No perinephric fat invasion or thrombus present. Surgical margins were negative.		
Reasoning Processes	Differential diagnosis: 1. Renal Cell Carcinoma (RCC) - RCC is the most common adult renal cancer and closely resembles nephroblastoma on imaging, presenting as a heterogeneous renal mass with possible lymph node involvement. However, histopathological analysis confirmed a triphasic nephroblastoma pattern and specific immunohistochemical markers, which ruled out RCC as a diagnosis. 2. Transitional Cell Carcinoma - Renal tumors involving the renal pelvis could suggest transitional cell carcinoma. However, imaging and histopathology findings localized the tumor strictly within the kidney and demonstrated features inconsistent with transitional cell carcinoma, excluding it as a possibility. 3. Adult Nephroblastoma/Wilms' Tumor - The patient's presentation, including a large renal mass, triphasic histopathologic pattern, and lymph node involvement, closely matched the typical features of rare adult nephroblastoma. Diagnosis was definitively confirmed based on the histological triphasic pattern and pathologic staging. 4. Other Rare Renal Neoplasms (e.g., Sarcoma, Lymphoma) - Rare possibilities such as sarcoma or lymphoma were considered based on the patient's age and presentation. However, histopathological findings definitively confirmed nephroblastoma, ruling out a mesenchymal or lymphoid origin.		
Diagnosis Results	Adult nephroblastoma (Wilms' tumor)		
Generation			
	Ground Truth	DeepSeek-R1	OpenAI-o3-mini
Diagnostic Decision-Making	Reasoning: <Step 1> Renal Cell Carcinoma (RCC) was considered because it is the most common adult renal cancer and closely resembles nephroblastoma on imaging, with a heterogeneous renal mass and possible lymph node involvement. <Step 2> Renal Cell Carcinoma (RCC) was excluded due to histopathological confirmation showing a triphasic nephroblastoma pattern and specific immunohistochemical markers consistent with nephroblastoma. <Step 3> Transitional Cell Carcinoma was considered because renal tumors involving the renal pelvis could suggest this diagnosis. <Step 4> Transitional Cell Carcinoma was excluded as imaging and histopathology localized the tumor strictly within the kidney and demonstrated features inconsistent with this carcinoma. <Step 5> Adult Nephroblastoma/Wilms' Tumor was considered due to the patient's presentation of a large renal mass, histopathologic triphasic pattern, and lymph node involvement, which matched typical features of adult nephroblastoma. <Step 6> Adult Nephroblastoma/Wilms' Tumor was confirmed based on the histological triphasic pattern and pathologic staging. <Step 7> Other rare renal neoplasms, such as sarcoma or lymphoma, were considered due to the age and presentation of the patient. <Step 8> Other rare renal neoplasms were excluded as histopathology definitively confirmed nephroblastoma rather than a mesenchymal or lymphoid origin. <Step 9> The diagnosis of adult nephroblastoma (Wilms' tumor) was confirmed post-nephrectomy by histopathology with the characteristic triphasic pattern, including blastemal, stromal, and epithelial components. <Step 10> The absence of fat invasion, negative surgical margins, and lymph node involvement supported the pathological staging as pT2b N2. Disease Diagnosis: Adult nephroblastoma (Wilms' tumor)	Reasoning: <step 1> The patient's persistent left abdominal pain, weight loss, fatigue, and imaging findings of a large heterogeneous renal mass with local extension and lymphadenopathy strongly suggest a malignant renal neoplasm. <step 2> The histopathologic confirmation of a triphasic pattern (blastemal, stromal, epithelial) is diagnostic of nephroblastoma (Wilms tumor), a tumor classically seen in children but rarely in adults. <step 3> Regional lymph node involvement (N2) and local extension (pT2b) confirm advanced disease without distant metastases. The absence of perinephric fat invasion or thrombus and negative surgical margins indicate localized resectability. <step 4> The combination of clinical, imaging, and histologic findings rules out other renal malignancies (e.g., renal cell carcinoma) and confirms nephroblastoma with metastatic lymph node involvement. Disease Diagnosis: Nephroblastoma (Wilms tumor) with regional lymph node metastasis (pT2b N2).	Reasoning: <step 1> The patient is a 48-year-old male presenting with persistent left abdominal pain, weight loss, and increasing fatigue over the past six months. These symptoms raise suspicion for an abdominal mass or a neoplastic process in the kidney or nearby organs. <step 2> The CT scan reveals a 12 cm heterogeneous mass in the left kidney with solid and cystic components, irregular borders, and regional lymphadenopathy. This suggests a malignant process such as a renal cell carcinoma, though a differential diagnosis must include other renal neoplasms. <step 3> The histopathology report confirms a nephroblastoma (Wilms' tumor) with a triphasic pattern involving blastemal, stromal, and epithelial components. Nephroblastoma is a type of kidney cancer primarily seen in children, but it can also occur in adults, though rarely. <step 4> The pT2b N2 staging indicates that the tumor is locally advanced (perinephric fat extension without invasion) and has regional lymph node involvement, but no distant metastases are observed. The patient's anemia and systemic symptoms are likely related to the cancer's local effects and possibly the regional metastasis. Disease Diagnosis: Nephroblastoma (Wilms' tumor), stage pT2b N2.
Metrics			
Accuracy	\	1	1
Efficiency	\	0.75	0.75
Factuality	\	1	1
Recall	\	0.8	0.8

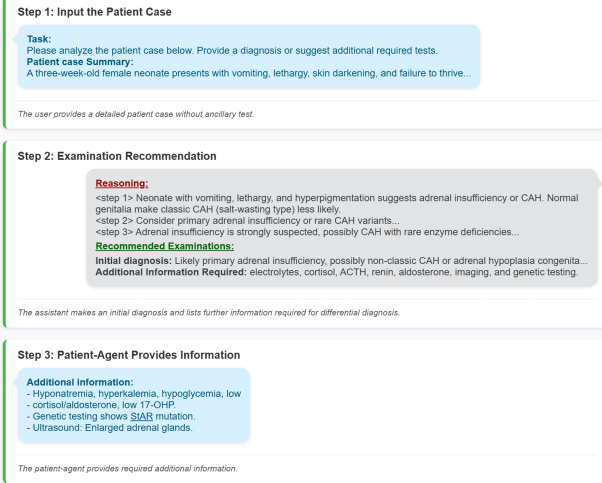
Supplementary Figure 6 | Case 6. A typical rare diseases case of diagnostic decision making, where exists specific tasks that are the golden standard for the confirmation of the diagnosis of the disease.

	Diagnostic Decision-making	Treatment Planning
Groudtruth answer	Metastatic pulmonary adenocarcinoma with EGFR L858R mutation, presenting as advanced non-small cell lung cancer (NSCLC) with brain and bone metastases (T4N3M1c staging).	Perform embolization of the type 2 endoleak using a micropuncture introducer set and a triple-coaxial system via the left femoral artery . Use a mixture of N-butyl-2-cyanoacrylate and iodized oil for embolization.
Predicted answer	Metastatic pulmonary adenocarcinoma (Stage IV, T4N3M1c) with *EGFR L858R* mutation, brain and bone metastases.	Repeat embolization of the left 3rd lumbar artery via the left deep iliac circumflex artery using a left femoral artery approach .
Judgement	Correct	Incorrect

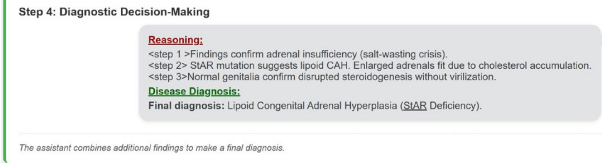
Supplementary Figure 7 | A case of comparison between diagnostic and treatment evaluation standards.

A.3 Supplementary Figures

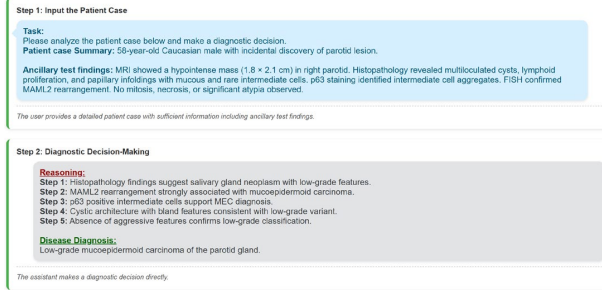
a. 1-turn Examination Recommendation



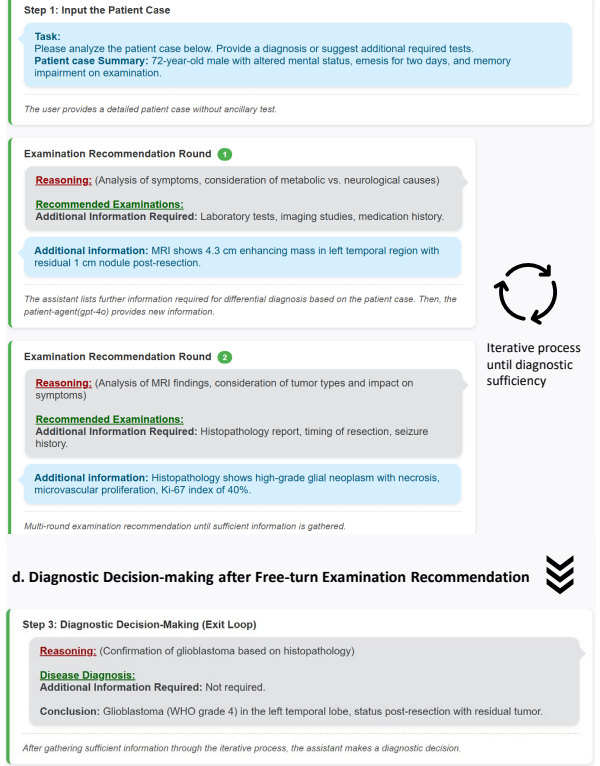
c. Diagnostic Decision-making after 1-turn Examination Recommendation



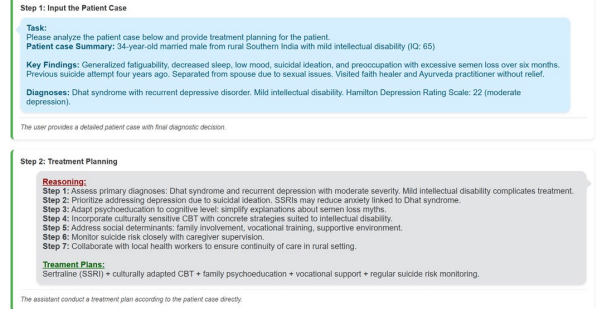
e. Oracle Diagnostic Decision-making



b. Free-turn Examination Recommendation



f. Treatment Planning



Supplementary Figure 8 | Overview of our evaluation settings. We consider three stages: examination recommendation, diagnosis decision-making, and treatment planning. **a**, **b** illustrate the 1-turn and free-turn interaction pipelines for examination recommendation. **c**, **d**, **e** depict the evaluation cases for diagnosis decision-making in 1-turn, free-turn, and oracle settings. Finally, **f** presents the treatment planning task in the oracle setting.

A.4 Supplementary Tables

Supplementary Table 1 | Results for recommended examinations on both “all diseases” and “rare diseases” in 1-turn and free-turn settings. The two-sided 0.95 confidence interval (CI) is reported in brackets.

Method	Model Size	All Diseases				Rare Diseases			
		1-turn		Free-turn		1-turn		Free-turn	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Close-source Reasoning LLMs									
OpenAI-o3-mini	–	33.75	38.47	33.57	38.22	33.73	38.25	32.94	37.24
		(32.21, 35.29)	(36.85, 40.10)	(31.96, 35.19)	(36.59, 39.85)	(31.60, 35.86)	(36.03, 40.47)	(30.77, 35.11)	(35.04, 39.45)
Gemini-2.0-FT	–	22.77	43.12	24.09	39.88	21.93	42.96	22.73	39.66
		(21.42, 24.12)	(41.41, 44.83)	(22.54, 25.63)	(38.21, 41.55)	(20.21, 23.64)	(40.56, 45.36)	(20.72, 24.75)	(37.30, 42.01)
Open-source Reasoning LLMs									
DeepSeek-R1	671B	31.50	43.61	32.23	40.67	30.92	43.73	31.64	40.18
		(29.97, 33.03)	(41.92, 45.30)	(30.58, 33.89)	(38.92, 42.42)	(28.81, 33.03)	(41.42, 46.04)	(29.36, 33.93)	(37.77, 42.59)
Qwen-QwQ	32B	24.43	39.90	25.39	40.83	24.08	39.44	25.47	40.79
		(23.19, 25.67)	(38.22, 41.57)	(24.16, 26.63)	(39.15, 42.51)	(22.36, 25.80)	(37.15, 41.74)	(23.76, 27.18)	(38.51, 43.08)
Baichuan-M1	14B	41.78	37.88	41.99	36.99	41.58	38.09	42.06	37.73
		(39.91, 43.64)	(36.31, 39.45)	(40.11, 43.87)	(35.40, 38.59)	(39.05, 44.12)	(36.01, 40.18)	(39.57, 44.56)	(35.58, 39.89)
DiagnoseGPT	34B	30.73	22.70	17.56	12.38	29.83	21.52	16.66	10.95
		(28.74, 32.73)	(21.14, 24.25)	(15.75, 19.37)	(11.04, 19.37)	(27.15, 32.52)	(19.40, 23.64)	(14.19, 19.14)	(9.27, 12.64)
MedGemma	27B	22.65	39.38	22.20	34.24	21.79	38.66	21.40	33.95
		(21.25, 24.04)	(37.62, 41.13)	(20.73, 23.68)	(32.43, 36.04)	(19.90, 23.67)	(36.22, 41.11)	(19.43, 23.37)	(31.49, 36.41)

Supplementary Table 2 | Results for examination recommendation reasoning on both “all diseases” and “rare diseases” in the 1-turn settings. For DeepSeek-R1, we assess its two types of reasoning, *i.e.*, those presented in the formal answer part and the default thinking part, recorded as “XX/xx” respectively. The two-sided 0.95 confidence interval (CI) is reported in brackets.

Method	Model Size	All Diseases		Rare Diseases	
		Efficiency	Factuality	Efficiency	Factuality
Close-source Reasoning LLMs					
OpenAI-o3-mini	–	95.17	95.89	92.73	95.90
		(94.47, 95.87)	(95.18, 96.61)	(91.16, 94.29)	(94.93, 96.87)
Gemini-2.0-FT	–	95.42	98.75	84.22	98.98
		(94.79, 96.05)	(98.38, 99.13)	(82.62, 85.83)	(98.52, 99.43)
Open-source Reasoning LLMs					
DeepSeek-R1	671B	98.59 /89.75	96.79 /95.38	95.96 /88.91	97.02 /95.10
		(98.17, 99.00)	(96.17, 97.41)	(94.83, 97.08)	(96.17, 97.87)
Qwen-QwQ	32B	86.53	94.87	76.34	94.53
		(85.34, 87.71)	(94.09, 95.64)	(74.02, 78.67)	(93.45, 95.62)
Baichuan-M1	14B	90.66	96.87	84.37	97.06
		(89.66, 91.66)	(96.35, 97.39)	(82.26, 86.48)	(96.36, 97.76)
DiagnoseGPT	34B	80.63	94.49	81.53	94.07
		(78.39, 82.87)	(93.35, 95.63)	(78.43, 84.62)	(92.33, 95.80)
MedGemma	27B	90.88	97.23	90.79	97.65
		(90.16, 91.60)	(96.72, 97.74)	(89.73, 91.85)	(96.99, 98.32)

Supplementary Table 3 | Results for disease diagnosis in the 1-turn, free-turn, and oracle evaluation settings on both “all diseases” and “rare diseases”. The two-sided 0.95 confidence interval (CI) is reported in brackets.

Method	Model Size	All Diseases			Rare Diseases		
		1-turn	Free-turn	Oracle	1-turn	Free-turn	Oracle
		Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
Human Reference							
5-year physician	-	-	-	76.67	-	-	73.33
Close-source Reasoning LLMs							
OpenAI-o3-mini	-	64.99	67.19	83.91	63.75	65.99	85.54
		(61.97, 68.02)	(64.21, 70.17)	(81.58, 86.24)	(59.48, 68.01)	(61.78, 70.19)	(82.43, 88.65)
Gemini-2.0-FT	-	68.55	68.76	86.83	66.40	68.43	88.60
		(65.60, 71.49)	(65.81, 71.70)	(84.69, 88.98)	(62.20, 70.59)	(64.31, 72.56)	(85.78, 91.41)
Open-source Reasoning LLMs							
DeepSeek-R1	671B	71.79	76.18	89.76	70.67	77.60	91.04
		(68.93, 74.64)	(73.47, 78.88)	(87.84, 91.68)	(66.63, 74.71)	(73.90, 81.30)	(88.51, 93.57)
Qwen-QwQ	32B	63.74	74.71	85.06	64.15	73.93	86.15
		(60.69, 66.79)	(71.95, 77.47)	(82.80, 87.32)	(59.90, 68.41)	(70.03, 77.83)	(83.09, 89.21)
Baichuan-M1	14B	66.77	61.23	84.43	67.21	60.69	88.39
		(63.78, 69.76)	(58.14, 64.33)	(82.13, 86.73)	(63.04, 71.38)	(56.36, 65.03)	(85.55, 91.23)
DiagnoseGPT	34B	54.44	39.60	79.62	52.34	35.44	79.63
		(51.28, 57.60)	(36.50, 42.71)	(77.07, 82.18)	(47.91, 56.78)	(31.19, 39.68)	(76.06, 83.21)
MedGemma	27B	61.23	64.23	84.22	60.69	65.17	86.35
		(58.14, 64.33)	(61.18, 67.27)	(81.91, 86.54)	(56.36, 65.03)	(60.94, 69.40)	(83.31, 89.40)

Supplementary Table 4 | Results for diagnostic decision-making reasoning in the 1-turn and oracle evaluation settings on “all diseases”. For **DeepSeek-R1**, we assess its two types of reasoning, *i.e.*, those presented in the formal answer part and the default thinking part, recorded as “XX/xx” respectively. The two-sided 0.95 confidence interval (CI) is reported in brackets.

Method	Model Size	1-turn reasoning			Oracle reasoning		
		Efficiency	Factuality	Completeness	Efficiency	Factuality	Completeness
Close-source Reasoning LLMs							
OpenAI-o3-mini	–	91.59	83.15	50.87	94.33	94.94	75.42
		(90.37, 92.81)	(81.29, 85.01)	(48.60, 53.13)	(93.61, 95.05)	(94.23, 95.64)	(74.14, 76.70)
Gemini-2.0-FT	–	83.77	87.17	54.45	95.89	98.23	83.28
		(82.56, 84.97)	(85.52, 88.83)	(52.20, 56.69)	(95.35, 96.43)	(97.84, 98.62)	(82.17, 84.40)
Open-source Reasoning LLMs							
DeepSeek-R1	671B	95.86/88.24	87.15/85.90	54.88/67.59	97.17/89.73	95.03/96.31	78.27/90.79
		(95.02, 96.71)	(85.54, 88.76)	(52.77, 56.99)	(96.65, 97.70)	(94.34, 95.71)	(77.17, 79.37)
Qwen-QwQ	32B	76.97	88.14	66.94	71.20	84.02	79.97
		(75.30, 78.64)	(86.48, 89.80)	(63.83, 70.04)	(69.46, 72.94)	(82.23, 85.80)	(78.11, 81.83)
Baichuan-M1	14B	82.91	88.62	53.43	92.80	96.84	75.11
		(81.34, 84.48)	(87.23, 90.01)	(51.17, 55.68)	(92.09, 93.51)	(96.35, 97.32)	(73.87, 76.36)
DiagnoseGPT	34B	80.99	89.14	25.44	90.42	93.72	79.20
		(78.78, 83.19)	(87.58, 90.70)	(23.39, 27.50)	(89.12, 91.73)	(92.87, 94.56)	(77.91, 80.48)
MedGemma	27B	90.22	88.23	49.91	96.78	95.96	87.72
		(89.42, 91.01)	(87.02, 89.44)	(47.61, 52.21)	(96.33, 97.24)	(95.41, 96.51)	(86.69, 88.74)

Supplementary Table 5 | Results for diagnostic decision-making reasoning in the 1-turn and oracle evaluation settings on “rare diseases”. For **DeepSeek-R1**, we assess its two types of reasoning, *i.e.*, those presented in the formal answer part and the default thinking part, recorded as “XX/xx” respectively. The two-sided 0.95 confidence interval (CI) is reported in brackets.

Method	Model Size	1-turn reasoning			Oracle reasoning		
		Efficiency	Factuality	Completeness	Efficiency	Factuality	Completeness
Close-source Reasoning LLMs							
OpenAI-o3-mini	–	92.73	81.28	49.83	94.80	95.02	76.44
		(91.16, 94.29)	(78.58, 83.99)	(46.59, 53.07)	(93.83, 95.77)	(94.14, 95.89)	(74.74, 78.14)
Gemini-2.0-FT	–	84.22	85.93	54.06	96.45	98.39	84.30
		(82.62, 85.83)	(83.53, 88.33)	(50.85, 57.28)	(95.74, 97.16)	(97.89, 98.89)	(82.75, 85.85)
Open-source Reasoning LLMs							
DeepSeek-R1	671B	95.96/88.91	84.76/84.70	54.10/67.54	97.61/90.33	94.75/96.83	79.00/91.14
		(94.83, 97.08)	(82.30, 87.22)	(51.05, 57.15)	(96.95, 98.27)	(93.77, 95.74)	(77.52, 80.48)
Qwen-QwQ	32B	76.34	89.05	66.53	72.25	84.30	80.70
		(74.02, 78.67)	(86.79, 91.31)	(62.14, 70.92)	(69.83, 74.68)	(81.83, 86.77)	(78.16, 83.24)
Baichuan-M1	14B	84.37	88.90	53.11	93.94	96.91	76.05
		(82.26, 86.48)	(87.06, 90.73)	(49.95, 56.27)	(93.07, 94.81)	(96.26, 97.56)	(74.31, 77.78)
DiagnoseGPT	34B	81.82	88.09	25.01	91.87	93.57	79.02
		(78.80, 84.85)	(85.85, 90.33)	(22.17, 27.85)	(90.18, 93.56)	(92.45, 94.69)	(77.20, 80.83)
MedGemma	27B	89.87	88.02	50.18	96.99	95.72	88.77
		(88.72, 91.02)	(86.31, 89.72)	(46.89, 53.47)	(96.41, 97.57)	(94.95, 96.48)	(87.46, 90.08)

Supplementary Table 6 | Results for treatment planning on both “all diseases” and “rare diseases”. We calculate LLM-powered reasoning metrics for both the final generation and reasoning processes. For **DeepSeek-R1**, we have assessed its two types of reasoning, *i.e.*, those presented in the formal answer part and the default thinking part, recorded as “XX/xx” respectively. The two-sided 0.95 confidence interval (CI) is reported in brackets.

Method	Model Size	All Diseases				Rare Diseases			
		Treatment Plans	Reasoning Processes			Treatment Plans	Reasoning Processes		
		Accuracy	Efficiency	Factuality	Completeness	Accuracy	Efficiency	Factuality	Completeness
Human Reference									
5-year physician	-	36.67	-	-	-	30.00	-	-	-
Close-source Reasoning LLMs									
OpenAI-o3-mini	-	27.03	94.67	96.77	70.63	23.17	95.06	96.81	69.86
		(23.09, 30.98)	(93.90, 95.44)	(96.06, 97.48)	(68.55, 72.71)	(16.65, 29.70)	(93.79, 96.33)	(95.51, 98.11)	(66.33, 73.38)
Gemini-2.0-FT	-	25.66	93.66	96.96	75.89	23.78	94.41	96.68	77.10
		(21.80, 29.52)	(92.82, 94.49)	(96.34, 97.59)	(73.81, 77.98)	(17.20, 30.37)	(93.07, 95.74)	(95.57, 97.78)	(73.68, 80.51)
Open-source Reasoning LLMs									
DeepSeek-R1	671B	30.51	95.25 /88.93	94.59 /95.93	68.08 /80.48	27.27	95.37 /89.19	95.22 /95.01	68.28 /81.05
		(26.43, 34.58)	(94.43, 96.08)	(93.63, 95.56)	(65.97, 70.20)	(20.41, 34.14)	(93.92, 96.82)	(93.62, 96.47)	(64.57, 71.99)
Qwen-QwQ	32B	20.89	84.76	94.40	77.66	19.61	83.31	94.05	78.74
		(17.12, 24.66)	(83.42, 86.10)	(93.44, 95.36)	(75.36, 79.96)	(13.25, 25.97)	(81.08, 85.54)	(92.00, 96.10)	(74.85, 82.63)
Baichuan-M1	14B	30.65	88.47	96.56	70.18	30.30	87.87	95.97	69.56
		(26.57, 34.72)	(87.37, 89.57)	(95.80, 97.31)	(68.04, 72.32)	(23.22, 37.39)	(85.96, 89.77)	(94.62, 97.31)	(65.91, 73.21)
DiagnoseGPT	34B	17.74	89.21	92.86	53.86	14.55	89.33	91.47	52.25
		(14.37, 21.12)	(86.87, 91.55)	(91.46, 94.26)	(50.80, 56.93)	(9.11, 19.98)	(85.61, 93.05)	(88.87, 94.06)	(46.91, 57.59)
MedGemma	27B	25.05	96.53	95.62	71.70	21.95	96.91	95.14	70.79
		(21.22, 28.88)	(95.84, 97.23)	(94.66, 96.59)	(68.91, 74.49)	(15.55, 28.35)	(95.77, 98.05)	(93.30, 96.98)	(65.76, 75.82)

Supplementary Table 7 | Comparison of Factuality Scores in Two Evaluation Rounds. “Original” refers to the results obtained during the submission in February 2025, while “Latest” represents the results recalculated four months later in June 2025. The "Consistency" line represents the proportion of reasoning steps that received the same judgment across the two metric calculation procedures.

Times	Date	Examination Recommendation	Diagnostic Decision-making		Treatment Planning
		1-turn	1-turn	oracle	
Original	2025-2	96.79	87.15	95.03	94.59
Latest	2025-6	97.55	87.77	95.03	95.60
Consistency(%)	–	97.13	93.20	96.20	95.30

Supplementary Table 8 | Comparison between LLMs and real physicians in evaluating final generation and reasoning processes. “Consistency” quantifies the alignment between LLM-based agentic evaluation results or individual physician results and the final manual majority rating results. “Precision” indicates the fraction of model-predicted positive cases that are true positives based on manual majority voting. “Recall” measures the fraction of actual positive cases that are successfully detected by the model.

Metric	Patient Stages	Agentic vs. Majority			Individual vs. Majority
		Precision (%)	Recall (%)	Consistency (%)	Consistency (%)
Accuracy	Diagnostic Decision-Making	97.78	95.65	94.00	92.00
	Treatment Planning	93.75	71.43	86.00	85.33
Efficiency	Diagnostic Decision-Making	97.67	97.67	95.56	94.83
	Treatment Planning	96.15	98.04	94.34	93.68
Completeness	Diagnostic Decision-Making	99.19	92.79	91.78	91.78
	Treatment Planning	99.01	84.75	86.33	84.67
Factuality	Diagnostic Decision-Making	98.84	98.84	97.75	96.24
	Treatment Planning	98.95	96.91	96.00	92.00

A.5 Prompt Collection

Prompt

Prompt 1. Prompt to classify case reports into "Diagnosis" or "Treatment Planning."

You are an experienced doctor. I will provide you with the title, abstract, and discussion of a case report. Please help me analyze whether the case report primarily focuses on the diagnostic/treatment process. If the report centers on diagnosis/treatment planning, please respond with "Yes." If the report does not primarily address the diagnostic/treatment planning task, please respond with "No."

Prompt

Prompt 2. Prompt for generating diagnosis data from case report.

As an experienced physician, you will receive a patient case report focused on diagnosis. Your task is to:

- Summarize the key information of the patient for diagnosis.
- Summarize the differential diagnosis process, including the rationale for each step and the reasons for considering or excluding specific diagnoses.
- Summarize the final diagnosis of the patient.

Ensure that your summaries are concise and accurate, based solely on the information provided in the case report. Please avoid referencing any images, tables, or other visual data (if any), as that data is no longer accessible.

If the case report is incomplete or does not meet the requirements for summarization, simply output: "I can't."

Format to follow:

Case Summary

Please provide a detailed medical history of the patient, including chief complaint, history of present illness, past medical history, family history, physical examination, results of ancillary tests, and other necessary information for the final diagnosis in the following format.

- Patient Information:
- Chief Complaint: If none, write "None."
- History of Present Illness: If none, write "None."
- Past Medical History: If none, write "None."
- Personal History: If none, write "None."
- Family History: If none, write "None."
- Physical Examination: If none, write "None."
- Ancillary Tests: If none, write "None."

Differential Diagnosis Process

- Firstly, based on the patient's chief complaint and clinical information, an initial list of potential diseases should be generated. This list will then be systematically evaluated by comparing each disease with the patient's symptoms, signs, and test results.
- Then the rationale for including or excluding each diagnosis will be explained based on clinical evidence, highlighting the selection of diagnostic tests that help confirm or rule out specific conditions. The process will involve step-by-step exclusion of less likely diagnoses, ultimately leading to the most probable diagnosis, which will be justified by the clinical reasoning and evidence supporting it.

Final Diagnosis (with explanation details)

Integrate the patient's clinical presentation, test results, and differential diagnosis process to summarize the final diagnosis. Briefly explain the basis for the diagnosis and highlight the key factors supporting this conclusion.

Diagnosis Results

Just Output the diagnostic result without any other explanation.

Prompt

Prompt 3. Prompt for generating treatment planning data from case report.

As an experienced physician, you will receive comprehensive patient information for treatment planning. Your task is to:

- Summarize provided clinical data, focusing on elements critical for treatment planning. Avoid any information after the treatment.
- Summarize the treatment plan for the patient without revealing any information about the treatment outcomes.
- Avoid referencing any images, tables, or other visual data (if any), as that data is no longer accessible. Ensure all conclusions are strictly based on the information provided. Do not infer or generate additional information beyond what is given. If the data is insufficient, output: "I can't."

Required Format:

Comprehensive Patient Profile

Organize medical information. Never involve any information after the treatment such as treatment results to avoid knowledge leakage. Also, avoid any information of treatment advice. Just state the patient case. The following is a template you can use to organize the information. You can add or remove content as needed.

- Patient Demographics: (Age/Sex/Other identifiers)
- Chief Complaint:
- History of Present Illness: (including current medication or surgical conditions ...)
- Past Medical History: (including surgical history, medication history ...)
- Allergies:
- Family History:
- Social History:
- Physical Exam:
- Lab Events:
- Imaging Events:
- Final Diagnostic Results:

Treatment Planning Analysis

Summarize the primary objectives of the treatment plan based on the patient's condition and needs. Then, summarize the comprehensive rationale for selecting the preferred treatment plan prior to implementation.

- Object: Objectives of the treatment plan
- Rationale: Combined with the Comprehensive Patient Profile above, explain the reasons for choosing this treatment plan

Selected Treatment for the Patient

Output the treatment in a sentence directly. Use imperative sentences and avoid past tense.

Prompt

Prompt 4. Prompt for body system classification

Please categorize the body parts involved in the health conditions and disease phenotypes discussed in the following case report. Use the provided title, abstract, and discussion sections to determine the most relevant category. If the body part is not listed, please output "Others." Output the category directly without any additional words. If there are multiple categories, please use \n to separate them. Categories:

Blood, Heart and Circulation
Bones, Joints and Muscles
Brain and Nerves
Digestive System
Ear, Nose and Throat
Endocrine System
Eyes and Vision
Immune System
Kidneys and Urinary System
Lungs and Breathing
Mouth and Teeth
Skin, Hair and Nails
Male Reproductive System

Prompt

Prompt 5. Prompt for disorder and condition classification

Please categorize the diseases and symptoms discussed in the following case report. Use the provided title, abstract, and discussion sections to determine the most relevant category from the list below. If the diseases and symptoms do not fit into any of the listed categories, please output "Others." Output the category directly without any additional words.

Categories:

Cancers
Diabetes Mellitus
Genetics/Birth Defects
Infections
Injuries and Wounds
Mental Health and Behavior
Metabolic Problems
Poisoning, Toxicology, Environmental Health
Pregnancy and Reproduction
Substance Use and Disorders

Prompt

Prompt 6. Prompt used to check whether a certain rare disease are mentioned in the patient case.

You are an experienced physician. You will be provided with the name of a rare disease, as well as the title, abstract, and Discussion section of a case report. Your task is to determine whether the case report is related to this rare disease. If it is related, output "YES" directly. If it is not related, output "NO" directly.

Prompt

Prompt 7. Prompt designed for GPT-4o to role-play as the patient and provide the required information.

You are a medical expert providing guidance to a junior physician on a patient case. The junior physician will ask you for additional diagnostic information based on the patient's case details and any available ancillary test results. Your role is to provide accurate and relevant responses regarding the availability of specific diagnostic information.

Guidelines:

1. You will receive the patient's case information and any relevant ancillary test results.
2. The junior physician will ask questions about additional diagnostic information needed for the case.
3. If there is relevant ancillary test information available for the requested diagnostic area, provide the details 4. accurately.
4. If there is no relevant ancillary test information available for the requested diagnostic area, simply state: "There is no relevant ancillary test information available for this request."

[Patient Case Summary without Ancillary Test]

[Ancillary Test Information]

Example Interaction:

Junior Physician: "Does the patient have any imaging studies like an X-ray or CT scan?"

Your Response:

If there is relevant imaging information available:

"Based on the available ancillary test results, the patient has undergone a chest X-ray which shows (specific findings)."

If there is no relevant imaging information available:

"There is no relevant ancillary test information available for this request."

Note: Your responses should be factual and based solely on the provided patient case information and ancillary test results. Avoid speculation or hypotheticals unless explicitly requested.

Prompt

Prompt 8. Prompt for asking information under 1-turn examination recommendation setting.

Please thoroughly examine the patient case summary presented below. Your objective is to perform a detailed diagnostic analysis utilizing all available information. Note that due to the potentially limited details, the preliminary diagnosis may encompass several possible conditions. Should you ascertain that the provided data is inadequate for a definitive conclusion, please enumerate any additional diagnostic tests or information that would be necessary. However, if you can deduce a conclusive diagnosis, please proceed to provide it. Too many requests for information are also inappropriate.

[Patient Case Summary without Ancillary Test]

Guidelines:

Evaluate the patient's symptoms, medical history, and all pertinent details from the case summary.

Formulate differential diagnoses based on your analysis.

If the information is not sufficient for a conclusive diagnosis, specify the further tests or details required.

Always following the response format in each turn of the dialogue:

Chain of Thought:

Please sort out your thinking process step by step, with each logical step in a separate paragraph, and use a format such as <step 1> to label each step.

<step 1> Specific thinking content of this step

<step 2> Specific thinking content of this step

...

<step n> Specific thinking content of this step

Conclusion:

Give a preliminary conclusion if possible, or summarize the current findings.

Additional Information Required:

[Indicate if further information is needed by specifying the required tests or data. If a conclusive diagnosis has been made and no additional information is necessary, only output "Not required." directly without any other words in this section.]

For example:

Not required.

or

1. Laboratory tests: details

2. Imaging: details

Prompt

Prompt 9. Prompt for making the diagnostic decision after 1-turn examination recommendation.

Please make a final diagnosis for the patient in light of the additional information provided below.

[Additional Information]

Guidelines:

- Evaluate the patient's symptoms, medical history, and all pertinent details from the case summary.
- Formulate differential diagnoses based on your analysis.

Always following the response format in each turn of the dialogue:

Chain of Thought:

Please sort out your thinking process step by step, with each logical step in a separate paragraph, and use a format such as <step 1> to label each step.

<step 1> Specific thinking content of this step

<step 2> Specific thinking content of this step

...

<step n> Specific thinking content of this step

Conclusion:

Directly output the diagnostic result without any other explanation.

Prompt

Prompt 10. Prompt for the first turn under the free-turn examination recommendation setting.

Please thoroughly examine the patient case summary presented below. Your objective is to perform a detailed diagnostic analysis utilizing all available information. Note that due to the potentially limited details, the preliminary diagnosis may encompass several possible conditions. Should you ascertain that the provided data is inadequate for a definitive conclusion, please enumerate any additional diagnostic tests or information that would be necessary. However, if you can deduce a conclusive diagnosis, please proceed to provide it. Too many requests for information are also inappropriate.

[Patient Case Summary without Ancillary Test]

Guidelines:

Evaluate the patient's symptoms, medical history, and all pertinent details from the case summary.

Formulate differential diagnoses based on your analysis.

If the information is not sufficient for a conclusive diagnosis, specify the further tests or details required.

Always following the response format in the following dialogue:

Chain of Thought:

Please sort out your thinking process step by step, with each logical step in a separate paragraph, and use a format such as <step 1> to label each step.

<step 1> Specific thinking content of this step

<step 2> Specific thinking content of this step

...

<step n> Specific thinking content of this step

Additional Information Required:

Indicate if further information is needed by specifying the required tests or data. If a conclusive diagnosis has been made and no additional information is necessary, only output "Not required." directly without any other words in this section.

For example:

Not required.

or

1. Laboratory tests: details

2. Imaging: details

Conclusion:

If do not require additional information, please provide a final conclusive diagnosis. Otherwise, summarize the current findings.

Prompt

Prompt 11. Prompt for subsequent turns under free-turn examination recommendation setting.

Here is the additional information you required. Please proceed with the analysis.

[Additional Information]

Always following the response format in each turn of the dialogue:

Chain of Thought:

Please sort out your thinking process step by step, with each logical step in a separate paragraph, and use a format such as <step 1> to label each step.

<step 1> Specific thinking content of this step

<step 2> Specific thinking content of this step

...

<step n> Specific thinking content of this step

Additional Information Required:

Indicate if further information is needed by specifying the required tests or data. If a conclusive diagnosis has been made and no additional information is necessary, only output "Not required." directly without any other words in this section.

For example:

Not required.

or

1. Laboratory tests: details

2. Imaging: details

Conclusion:

If do not require additional information, please provide a final conclusive diagnosis. Otherwise, summarize the current findings.

Prompt

Prompt 12. Prompt for instructing LLMs to diagnose based on patient case.

Please carefully study the following patient case summary, conduct a comprehensive and in-depth diagnostic analysis, and clearly provide the final diagnosis result.

[Patient Case Summary]

Format to Follow:

Reasoning:

Please sort out your thinking process step by step, with each logical step in a separate paragraph.

<step 1> Specific thinking content of this step

<step 2> Specific thinking content of this step

...

<step n> Specific thinking content of this step

Answer:

Just output the diagnostic result without any other explanation.

Prompt

Prompt 13. Prompt for instructing LLMs to devise a treatment plan based on a patient case.

Please carefully study the following patient case summary, conduct a comprehensive and in-depth treatment planning analysis, and clearly provide the selected treatment for the patient.

[Patient Case Summary]

Format to Follow:

Chain of Thought:

Please sort out your thinking process step by step, with each logical step in a separate paragraph, and use a format such as <step 1> to label each step.

<step 1> Specific thinking content of this step

<step 2> Specific thinking content of this step

...

<step n> Specific thinking content of this step

Answer:

Just output the selected treatment for the patient without any other explanation.

Prompt

Prompt 14. Prompt for reformatting the unstructured ground-truth rationale.

Task Overview

Given a medical case analysis problem with multiple reasoning steps [Text to be Organized], reorganize it into clearly structured steps, separated by newline characters.

Organization Requirements

Convert the original solution into a clear, structured reasoning process, while ensuring:

- All original key information is preserved, but if multiple sentences discuss the same topic or serve the same logical reasoning purpose, they can be combined into one reasoning step.
- No new explanations or reasoning are added.
- No steps are omitted.

Requirements

- Each step must be atomic (one conclusion per step).
- There should be no content repetition between steps.
- The final answer determination is also considered a step in the logical reasoning.

Output Requirements

1. There should be no newline characters within each step, and each step should be separated by a single newline character.
2. For highly repetitive reasoning patterns, output them as a single step.
3. Output a maximum of 10 steps.

Output Format

<Step 1> Content of this reasoning step...

<Step 2> Content of this reasoning step...

...

<Step n> Content of this reasoning step...

Below is the text that needs to be reorganized into reasoning steps:

[Text to be Organized]

Prompt

Prompt 15. Prompt for efficiency classification of each reasoning step.

Task Description

Please analyze and determine the type of the current thinking step based on the provided medical case analysis process, all previous thinking steps, the known patient medical record, and the final reasoning goal. The current thinking step should be classified into one of the following types:

- 1.Citation: A simple citation, summary, or restatement of information from the patient's medical record, without generating new reasoning or conclusions.
- 2.Repetition: A repetition of previous thinking processes, without providing new information or advancing the reasoning process.
- 3.Reasoning: Providing information beyond what is known, or deriving new conclusions from known information, or proposing new possibilities, which moves the thinking process towards the correct answer and has a direct or indirect effect on the final reasoning goal.
- 4.Redundancy: Providing new information or possibilities that do not help in reaching the final answer and do not advance the reasoning process.

Note

When determining the type, ensure to fully consider the logical relationship and reasoning process between the current thinking step, previous thinking steps, the patient's medical record, and the reasoning goal. If the current thinking step corresponds to multiple types, select the most appropriate one based on its contribution to the reasoning goal. Maintain objectivity and accuracy in judgment, avoiding subjective assumptions.

Output Requirements

Only output your classification of the current thinking step, with possible values being "Citation|Repetition|Reasoning|Redundancy". Do not output any other content.

Output Format

Choose one from "Citation", "Repetition", "Reasoning", and "Redundancy".

Now, please classify the following input based on the instructions above:

[Current Thinking Step]

[All Previous Thinking Steps]

[Known Patient Medical Record]

[Final Reasoning Goal]

Prompt

Prompt 16. Prompt to extract keywords from the medical text for search engines.

Task Description

You will receive a medical analysis text description [Text Description to be Judged], which involves the analysis and reasoning of a patient's condition. Your task is to deeply analyze this description, judge the correctness of the medical knowledge described, and the key medical facts supporting the logical reasoning. For this, you need to determine the key medical knowledge points required to make this judgment and provide keywords for retrieving these knowledge points. Ensure your analysis is comprehensive, accurate, and covers all key information in the text description. Please follow these steps:

1. Carefully read and understand the provided medical analysis text description.
2. Analyze the medical concepts, symptoms, diagnostic methods, treatment principles, etc., involved in the text.
3. Identify the logical reasoning steps in the text and judge whether they are based on correct medical knowledge.
4. Determine the key medical knowledge points required to verify the correctness of these logical reasoning steps.
5. Extract keywords for retrieving these medical knowledge points.

Output Requirements

Directly output the list of required keywords, separated by commas, with no other content.

Format to Follow

Keyword1, Keyword2,...

Below is the text description to be judged:

[Text Description to be Judged]

Prompt

Prompt 17. Prompt to verify whether each step adheres to medical priors.

Task Description

You will receive three types of content:

- Patient Case Summary: A summary containing key information about the patient, including basic information, medical history, examination results, etc.
- Text Description to be Judged: A medical text description regarding the patient's condition, diagnosis, or treatment, which you need to judge for its medical knowledge.
- Known Correct Information: Verified correct medical knowledge related to the above text description, serving as a reference for your judgment.

Your task is to deeply analyze the provided text description[Text Description to be Judged], judge whether its description of the involved medical knowledge is accurate, and whether the key medical facts supporting the logical reasoning are correct. Please proceed as follows:

- 1.Judge the correctness of the description based on the [Patient Case Summary], [Known Correct Information], and your medical knowledge;
- 2.If you are still uncertain about a particular description, list search keywords specific to that description. The listed search keywords should be targeted and accurate, helping to further verify the uncertain description.

Output Requirements

- 1.The output format must be valid JSON format, with no other content.
- 2.Output your judgment in the judgment field, with optional values "Correct|Wrong|Search".
- 3.If further search is needed, list the search keywords for the questions you are uncertain about in the keywords__to__search field. If the judgment is "Correct" or "Wrong", this field should be "None".

Format to Follow:

```
{{  
"judgment": "Correct|Wrong|Search",  
"keywords__to__search": "None|keywords",  
}}
```

Below is the patient information, known correct information, and the text description that needs to be judged for the patient:

[Patient Case Summary]

[Known Correct Information]

[Text Description to be Judged]

Prompt

Prompt 18. Prompt to verify if the provided reasoning step is included in the generated reasoning process.

Task Overview

Given a reasoning step from a medical case analysis problem, determine whether it appears or is covered in the reference reasoning process.

Task Requirements

Evaluate whether the content of the step to be analyzed is the same as or related to any step in the reference reasoning process. Or whether the reference reasoning process covers the core meaning or logic expressed in the reasoning step to be analyzed.

Output Requirements

Only output your judgment result on the [Reasoning Step to be Analyzed], with optional values “Yes|No”, do not output any other content.

Output Format

[Yes|No]

Below is the reasoning step to be analyzed and the reference reasoning process:

[Reasoning Step to be Analyzed]

[Reference Reasoning Process]

Prompt

Prompt 19. Prompt for final diagnosis accuracy

You are a professional medical diagnosis evaluation system. Now, you will receive two diagnosis results: one is the diagnosis predicted by the model, and the other is the verified correct diagnosis . Your task is to judge whether the model-predicted diagnosis is correct.

When evaluating, please consider the following factors:

- 1.The same disease may have multiple aliases, for example, “Heart disease” may also be called “Cardiac disease”.
- 2.There may be diversity in language expression, for example, “heart attack” and “myocardial infarction” may refer to the same disease.
- 3.Only judge whether the diagnosis result is correct, information such as the cause of the disease, symptoms, and treatment recommendations are not included in the evaluation scope.
- 4.If the correct diagnosis is included in the predicted diagnosis but some additional complications are mentioned, it is also considered correct.

Output Requirements

Only output your judgment result on the model-predicted [Diagnostic Decision Made by the LLMs] as “Correct|Wrong”, do not output any other content.

Format to Follow:

[Correct|Wrong]

Below is the diagnosis result predicted by the model and the correct diagnosis:

[Diagnostic Decision Made by the LLMs]

[Ground-truth Diagnostic Decision]

Prompt

Prompt 20. Prompt for treatment plan accuracy

Task Description

As a professional medical treatment planning evaluation system, you will now receive two treatment plan results for assessment: one is the treatment plan predicted by the model, and the other is the verified correct treatment plan. Your task is to determine whether the model-predicted treatment is accurate.

When evaluating, please consider the following factors:

1. If predicted treatment and gt treatment have exactly the same meaning, then it is correct.
2. If the correct treatment plan is included in the predicted treatment but some additional care are mentioned, it is also considered correct
3. Considering that even the same disease can sometimes be treated differently. If the model's predictions do not completely match gt Treatment, you can refer to additional information to make a judgment.
4. If the predicted treatment and the ground-truth treatment do not convey the same meaning, and there is no supporting evidence in the additional information to suggest that the predicted treatment is also applicable to the disease, it is considered wrong.

Output Requirements

Only output your judgment result on the model-predicted as "Correct|Wrong", do not output any other content.

Format to Follow:

[Correct|Wrong]

Below is the result predicted by the model and the correct Treatment plan:

[Predicted Treatment]

[Ground-truth Treatment]

[Additional Information]

Prompt

Prompt 21. Prompt to reformat additional information of auxiliary examinations into structured format.

Task Overview

You will receive an output from an auxiliary diagnostic and treatment large model detailing the additional information required for patient diagnosis analysis [Raw Output Text to be Organized]. Your task is to categorize this content into different information requirement categories, and output it in a JSON dictionary format.

Organization Requirements

1. Clearly categorize the original required information output into different information need categories.
2. Each point should clearly indicate:
 - "type": The major category of the test item (e.g., laboratory tests, imaging examinations, medical history inquiries, etc.)
 - "test_name": The specific name of the test item (e.g., MRI scan, CT scan, thyroid function test, lump biopsy, etc.)
 - "info_required": The specific content or purpose of the requirement (e.g., to rule out malignant tumors, to better observe the tympanic membrane and middle ear structure, to assess retinal function, etc.)
3. Retain all original content, but if multiple checks are for obtaining the same information or have the same purpose, they can be combined into one point, but the "test_name" field should include all these checks.
4. Do not add new required content; all information should originate from the original output [Raw Output Text to be Organized].
5. Do not omit any steps.

Output Requirements

1. The output format must be a valid JSON format without any other content.
2. Output the type of the test item in the "type" field; output the specific name of the test item in the "test_name" field; output the specific content or purpose of the required information in the "info_required" field.

Output Format

```
“json
{{
"type": "Major Category of the Test Item",
"test_name": "Specific Name of the Test Item",
"info_required": "Specific Information Required or the Purpose of the Test"
}}
```

Below is the raw output text that needs to be reorganized:

[Raw Output Text to be Organized]