MedProbeBench: Probing Evidence-Grounded Medical Deep Research

Anonymous ACL submission

Abstract

Recent advances in deep research (DR) systems enable large language models to retrieve, synthesize, and reason over large-scale external knowledge. However, existing medical DR benchmarks predominantly rely on simple multi-hop retrieval questions, lacking objective, fine-grained assessment of evidence-grounded expert documentation. To address this gap, we introduce **MedProbeBench**, a benchmark leveraging authoritative medical guidelines as expert-level references. These guidelines impose stringent demands on neutrality, comprehensiveness, evidence grounding, terminological precision, and citation accountability, posing substantial challenges for DR agents. We further propose **MedProbe-Eval**, an evaluation framework with: *(1) Holistic Rubrics* (1,200+ task-adaptive criteria) and *(2) Fine-grained Evidence Verification* grounded in 5,130+ atomic claims. Evaluation of 17 LLMs and DR agents reveals critical gaps in evidence integration, highlighting the distance between current capabilities and evidence-grounded medical documentation. Project & Dataset: `https://anonymous.4open.science/r/MedProbeBench-8742`

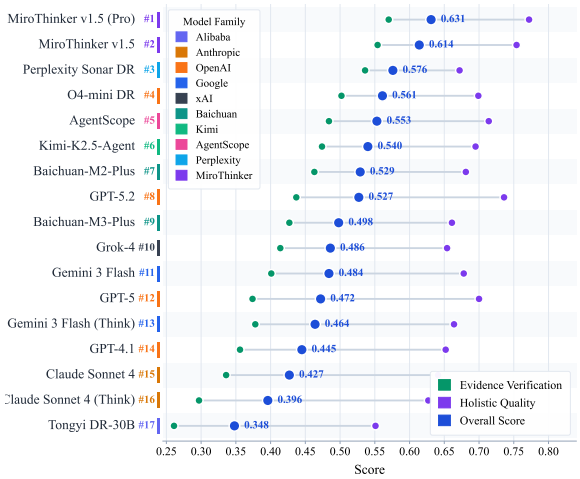

Figure 1: Performance of 17 LLMs and deep research agents on MedProbeBench. Even the best-performing system leaves substantial room for improvement.

1 Introduction

Deep research (DR) agents (Du et al., 2025) are designed to autonomously plan, search, and synthesize external evidence into structured knowledge artifacts. Recent medical AI systems have shown growing potential in clinical decision support (Yang et al., 2025), diagnostic reasoning (Wang et al., 2025a; Su et al., 2025), and broad biomedical synthesis (Huang et al., 2025). However, evaluating medical DR requires tasks that go beyond retrieving correct answers. One such task is the generation of *expert-level medical documentation*, where systems must produce comprehensive, terminologically precise, and professionally structured clinical content with verifiable citations for key claims. Clinical guidelines provide a natural testbed for this capability, because they require coherent synthesis of evidence across multiple clinical dimensions together with explicit citation accountability. Unlike question answering, this task stresses the full deep-research stack at once: planning a structured document, reconciling heterogeneous primary sources, and attaching each key claim to a traceable reference under standardized clinical headings.

Despite this need, existing benchmarks do not yet evaluate medical DR at the level required by expert documentation. *First, medical DR benchmarks remain centered on answer finding rather than document synthesis.* MedBrowseComp (Chen et al., 2025) focuses on multi-hop fact retrieval and short-answer formats, while MIRAGE (Xiong et al., 2024) targets retrieval-augmented QA. These settings test whether systems can locate isolated medical facts, but not whether they can synthesize evidence across sections, clinical dimensions, and

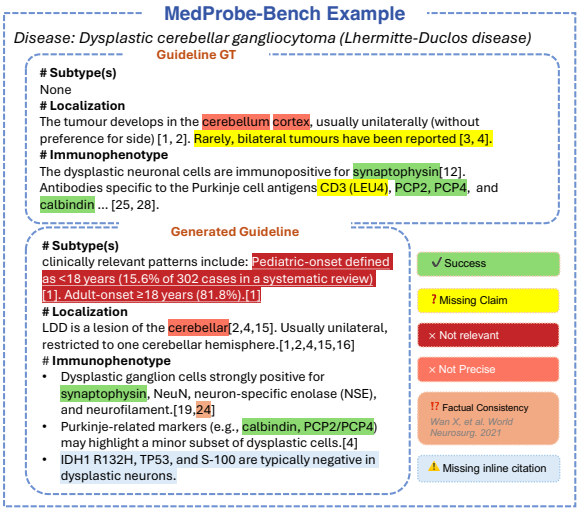

Figure 2: A characteristic failure mode in AI-generated clinical guidelines by DR agent.

sources into a coherent expert document. They therefore miss failures that arise only at document scale, such as missing diagnostic subsections, conflated molecular subtypes, or evidence cited at the wrong granularity, none of which can be exposed by multi-hop QA scoring. *Second, broader DR benchmarks lack the domain-specific accountability required in medicine.* DeepResearch Bench (Du et al., 2025) and related general benchmarks evaluate autonomous retrieval, planning, and long-horizon synthesis, but they typically rely on general-domain tasks and holistic quality judgments. Expert medical documentation requires stronger standards: outputs must satisfy clinically grounded criteria such as comprehensiveness, terminological precision, taxonomy adherence, and practical utility, while each key claim must be checked against its cited evidence. Without such claim-level verification, evaluation cannot distinguish genuinely evidence-grounded synthesis from fluent text with superficial citation support, as illustrated in Figure 2. The most common medical failure mode is precisely this: citations that are present and on-topic but do not actually support the specific quantitative or causal statement they accompany.

To address these gaps, we introduce **MedProbeBench**, a benchmark that operationalizes medical deep research as the generation of evidence-grounded expert medical documentation. We use physician-validated authoritative clinical guidelines as expert references, because such guidelines naturally impose stringent requirements on neutrality, comprehensiveness, structural clarity, terminological precision, and citation accountability for key medical claims. Specifically, we build on the WHO Classification of Tumours (5th Edition), whose fixed clinical schema, primary-literature citations, and coverage of both common and recently reclassified rare subtypes make it well suited to probing evidence discipline rather than only surface form.

Our MedProbeBench comprises 50 physician-validated clinical guidelines spanning 5 medical specialties, with 5,133 atomic claims (of which 2,203 carry verifiable reference anchors). For evaluation, we propose **MedProbe-Eval**, a dual-tier assessment framework featuring: *(1) Holistic Rubrics* with 1,236 task-adaptive criteria across four clinically grounded dimensions to assess overall guideline quality, and *(2) Fine-grained Evidence Verification* with three-dimensional reliability metrics to verify claim-level evidence grounding. The two tiers are complementary: the holistic tier captures document-level quality, while evidence verification isolates cases where surface quality is preserved but the underlying citation chain is broken.

As shown in Figure 1, our evaluation of 17 state-of-the-art systems reveals that even the best-performing agent leaves substantial room for improvement, highlighting a significant gap toward expert-level guideline generation. The gap is concentrated on reasoning-intensive content: average accuracy is only 28.8% on mechanistic, 22.2% on differential, and 19.2% on therapeutic claims, even when the surface text is fluent and well-structured. The contributions of this work are threefold:

1) We curate MedProbe-Dataset, a dataset of physician-validated clinical guidelines with claim-level annotations, designed to evaluate deep research in expert-level medical synthesis. 2) We propose MedProbe-Eval, a dual-tier evaluation framework combining holistic rubric-based quality assessment with fine-grained evidence verification, enabling systematic and reproducible benchmarking of medical deep research agents. 3) We conduct extensive experiments across 17 diverse systems to reveal persistent gaps among current DR agents.

2 Related Work

2.1 Medical LLM Benchmarks

Early medical LLM benchmarks predominantly evaluate closed-book or short-form question answering, exemplified by MedQA (Jin et al., 2021),

MMLU-Medical (Hendrycks et al., 2020), MedQ-Bench (Liu et al., 2025) and PubMedQA (Jin et al., 2019). While effective for assessing factual knowledge, these benchmarks emphasize isolated recall. A recent systematic review (Gong et al., 2025) identifies a core limitation: most evaluations assess knowledge in isolation rather than practice-oriented workflows that reflect real clinical expertise. Subsequent benchmarks introduce greater task complexity or retrieval augmentation (Zuo et al., 2025; Arora et al., 2025; Ning et al., 2026; Xiong et al., 2024), yet continue to focus on answering questions with definitive answers. Several recent efforts move toward guideline-aware or evidence-grounded evaluation. AMEGA (Fast et al., 2024) evaluates LLM adherence to medical guidelines through open-ended clinical scenarios across 13 specialties, but assesses compliance with existing guidelines rather than the ability to synthesize and generate them. MedGUIDE (Li et al., 2025) benchmarks clinical decision-making against 55 NCCN oncology decision trees using multiple-choice questions, focusing on protocol adherence rather than open-ended document generation. In contrast, MedProbeBench directly targets deep research by operationalizing it through the generation of evidence-grounded expert medical documentation.

2.2 Deep Research Systems and Benchmarks

Deep research systems support multi-step information gathering and synthesis beyond single-turn question answering. Early work such as WebGPT (Nakano et al., 2021), together with subsequent commercial systems including Perplexity Deep Research (Perplexity AI, 2025), OpenAI Deep Research (OpenAI, 2024), and Gemini Deep Research (DeepMind, 2024), demonstrates the effectiveness of iterative retrieval and long-form synthesis. Agent-based medical systems such as AURA (Fathi et al., 2025), MedAgent-Pro (Wang et al., 2025b), and TxAgent (Gao et al., 2025) further demonstrate advanced multi-step reasoning and tool use in clinical domains, while Tongyi DeepResearch (Team et al., 2025b) and MiroThinker (Team et al., 2025a) extend these capabilities through longer contexts and interactive reasoning. Benchmarks such as ResearchRubrics (Sharma et al., 2025), DeepResearch Bench (Du et al., 2025), and BrowseComp (Wei et al., 2025) aim to standardize the evaluation of deep research agents, but remain largely domain-agnostic and lack medical-specific evaluation criteria. MedProbeBench bridges this gap by targeting medical deep research agents with evidence-grounded expert medical documentation as the evaluation task.

3 Constructing MedProbeBench

3.1 Task Definition

MedProbeBench evaluates medical deep research through evidence-grounded guideline generation. Given a disease specification $\mathcal{D}$ and a required section schema $\mathcal{S} = \{s_1, \ldots, s_n\}$, a system must produce a guideline $\hat{\mathcal{G}}$ that is structurally complete, terminologically precise, and grounded in verifiable external evidence. This formulation turns deep research into a document-level task: the system must not only retrieve relevant sources, but also synthesize them into clinically organized content and support key claims with traceable citations.

We define key claims as atomic clinical statements spanning seven knowledge types: Factual, Mechanistic, Clinical, Diagnostic, Differential, Prognostic, and Therapeutic. This claim-level representation connects the construction of MedProbeBench to the fine-grained evidence verification introduced in Section 4.

3.2 Gold Standard Collection

We curate gold-standard guidelines from the WHO Classification of Tumours (5th Edition), a widely adopted clinical reference standard in oncology. MedProbeBench covers 50 physician-validated disease entities from five representative volumes: Central Nervous System Tumours (Louis et al., 2021), Soft Tissue and Bone Tumours (of Tumours Editorial Board et al., 2020), Digestive System Tumours (Nagtegaal et al., 2019), Thoracic Tumours (Tsao et al., 2022), and Haematolymphoid Tumours (Li et al., 2024). These sources provide diverse diagnostic criteria, molecular mechanisms, and disease-specific evidence patterns, making them suitable for benchmarking long-form medical synthesis.

3.3 Shared Structuring and Dataset Statistics

To compare expert guidelines and model outputs fairly, we convert all documents into a shared section- and claim-level representation, as shown in Figure 3. Gold-standard sources are first converted from PDF or HTML into structured Markdown with MinerU (Wang et al., 2024) and reviewed by board-certified medical professionals. Both gold

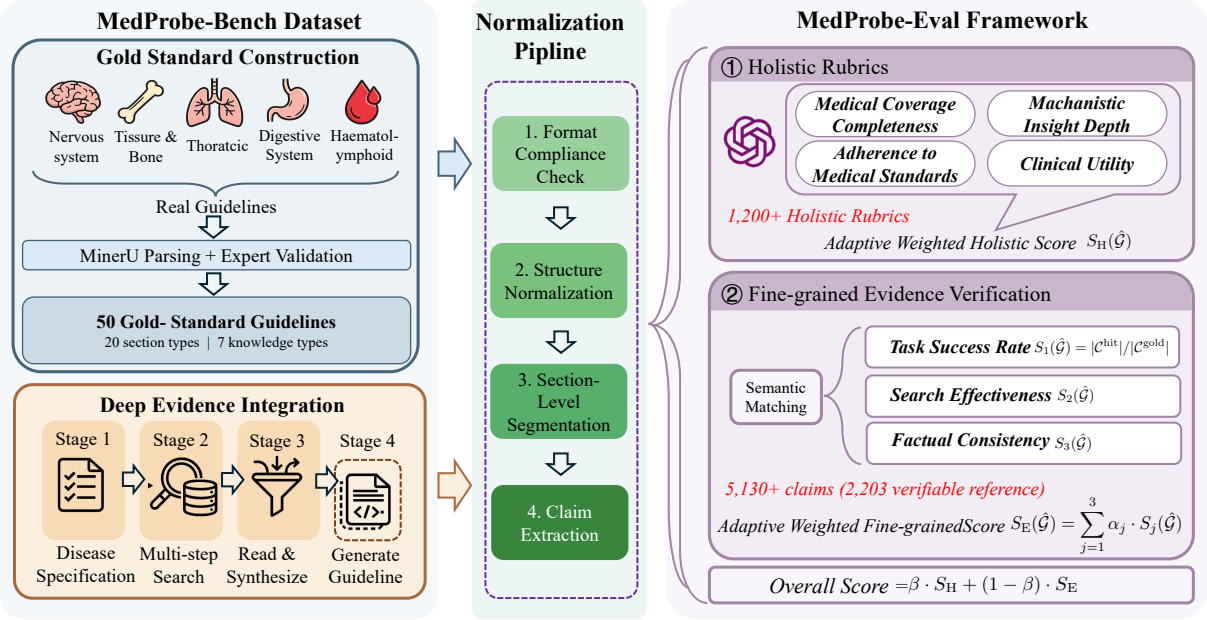

Figure 3: **MedProbeBench Framework.** A unified pipeline that normalizes and structures both clinical guidelines and generated reports into section- and claim-level representations, enabling fine-grained evidence-grounded evaluation.

and generated documents then pass through the same five-stage pipeline: format compliance checking, Markdown structure normalization, section-level segmentation, atomic claim extraction with in-text citation markers, and citation normalization into canonical PubMed- or DOI-based references when possible.

This shared representation reduces formatting bias and enables direct comparison between curated guidelines and generated reports. Overall, MedProbeBench contains 50 tasks, with 9–12 disease entities per domain, and 5,133 atomic claims across 20 guideline section types (e.g., Essential Diagnostic Criteria, Diagnostic Molecular Pathology, Histopathology, Immunophenotype, Prognosis and Prediction, Differential Diagnosis) and seven knowledge types (Factual, Mechanistic, Clinical, Diagnostic, Differential, Prognostic, Therapeutic). Among them, **2,203 (42.92%)** claims carry at least one supporting citation, providing verifiable reference anchors for fine-grained evidence verification. Per-domain statistics are summarized in Figure 4; the complete section list and citation coverage are reported in Appendix E and Appendix E.4. The resulting claim distribution is intentionally heterogeneous: factual claims account for 47.18% of the benchmark, while diagnostic (20.65%), mechanistic (9.39%), prognostic (8.63%), clinical (7.38%), differential (4.81%),

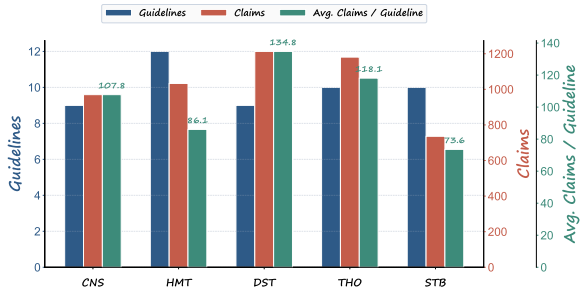

Figure 4: Dataset overview of MedProbeBench across medical domains. CNS: Central Nervous System Tumours; STB: Soft Tissue and Bone Tumours; DST: Digestive System Tumours; THO: Thoracic Tumours; HMT: Haematolymphoid Tumours.

and therapeutic (1.95%) claims together constitute more than half of all claims. This composition ensures that MedProbeBench tests not only descriptive recall, but also the reasoning-oriented evidence synthesis required by expert clinical documentation.

4 Evaluation Framework

MedProbe-Eval assesses generated guidelines from two complementary perspectives: *holistic quality*, which measures whether the document meets expert clinical standards, and *fine-grained evidence verification*, which tests whether generated claims are retrieved, cited, and supported at the claim

level. This design separates surface-level document quality from evidence fidelity, allowing us to identify fluent but weakly grounded medical synthesis. Evaluation prompts and implementation details are provided in Appendix A.

4.1 Holistic Rubrics

We evaluate each generated guideline $\hat{\mathcal{G}}$ along four clinically grounded dimensions: *Medical Coverage Completeness*, *Mechanistic Insight Depth*, *Adherence to Medical Standards*, and *Clinical Utility*. These dimensions capture whether the document is comprehensive, mechanistically informative, consistent with authoritative terminology and classification systems, and useful for clinical reasoning.

To reduce calibration drift in LLM-based scoring, we use a strict GT-anchored protocol. The reference guideline is treated as the expert-level anchor, and the generated guideline is scored by its deviation from that reference rather than by free-form preference. Missing content, weaker specificity, vague terminology, or unsupported substitutions reduce the score, while identical section titles alone are not considered evidence of substantive equivalence. This protocol is especially important for medical documentation, where a generated guideline can preserve the expected structure while omitting clinically decisive details.

The holistic score is computed as a task-adaptive weighted average:

$$S_{\mathrm{H}}(\hat{\mathcal{G}}) = \sum_{i=1}^{4} w_i \cdot s_i(\hat{\mathcal{G}}), \text{where} \sum_{i=1}^{4} w_i = 1 \quad (1)$$

Each dimension score $s_i(\hat{\mathcal{G}})$ is assigned by an LLM-based evaluator on a 0–10 scale. Following DeepResearch Bench (Du et al., 2025), weights w_i are derived from the task specification and section schema, then fixed for all systems on the same task (Appendix E.4).

4.2 Fine-grained Evidence Verification

We decompose each gold and generated guideline into atomic claims using $\mathcal{D} : \mathcal{G} \rightarrow \mathcal{C}$, where each claim is represented as $c_i = (t_i, \mathcal{R}_i, \tau_i)$: claim text, associated cited references, and clinical claim type. Let $\mathcal{C}^{\text{gold}} = \mathcal{D}(\mathcal{G}^{\text{gold}})$ and $\mathcal{C}^{\text{gen}} = \mathcal{D}(\hat{\mathcal{G}})$ denote the gold and generated claim sets. Generated claims are matched to gold claims using a two-stage semantic matching procedure with claim-level deduplication, producing the matched set $\mathcal{C}^{\text{hit}}$ (Appendix A.2.3, Appendix A.2.2).

We define three evidence-verification metrics. **(1) Task Success Rate** measures the proportion of gold claims recovered by the generated guideline:

$$S_1(\hat{\mathcal{G}}) = \frac{|\mathcal{C}^{\text{hit}}|}{|\mathcal{C}^{\text{gold}}|} \quad (2)$$

(2) Search Effectiveness evaluates whether generated references recover the evidence used by the gold guideline. Let Recall denote section-level reference recall after content-equivalence matching, and let $\mathrm{Quantity}$ be a bounded score for the amount of retrieved evidence. We compute:

$$S_2(\hat{\mathcal{G}}) = 0.6 \cdot \mathrm{Recall} + 0.4 \cdot \mathrm{Quantity}. \quad (3)$$

Detailed matching and quantity definitions are provided in Appendix A.2.5.

(3) Factual Consistency checks whether generated claims are supported by accessible cited content. For generated claims with at least one resolvable reference URL, we retrieve the cited sources and use an LLM-based verifier $\mathrm{verify}(c)$ to judge claim support. The score is:

$$S_3(\hat{\mathcal{G}}) = \frac{1}{|\mathcal{C}^{\text{url}}|} \sum_{c \in \mathcal{C}^{\text{url}}} \mathrm{verify}(c), \quad (4)$$

where $\mathcal{C}^{\text{url}}$ denotes generated claims with resolvable references. Verification details are provided in Appendix A.2.6.

The three metrics are combined into an overall evidence verification score:

$$S_{\mathrm{E}}(\hat{\mathcal{G}}) = \sum_{j=1}^{3} \alpha_j \cdot S_j(\hat{\mathcal{G}}), \text{where} \sum_{j=1}^{3} \alpha_j = 1 \quad (5)$$

As with holistic scoring, weights α_j are task-adaptive and fixed across systems for the same task (Appendix E.4). Figure 2 illustrates how this claim-level evaluation can reveal evidence-grounding failures that are not visible from surface fluency alone.

5 Performance Evaluation

5.1 Experimental Setup

Evaluated Systems. We evaluate 17 systems across three categories: (1) *LLMs with Search Tools*, including Claude Sonnet 4 (Anthropic, 2025), Gemini 3 Flash Preview (Google, 2025), GPT-4.1 (OpenAI, 2025b), GPT-5 (OpenAI, 2025d), GPT-5.2 (OpenAI, 2025e), Grok-4 (xAI, 2025), Baichuan-M2-Plus (Baichuan Intelligence, 2025), and Baichuan-M3-Plus (Baichuan Intelligence, 2026); (2) *Commercial Deep Research*

Model	Overall	Holistic Quality					Fine-grained Evidence Verification			
	Score↑	Compl.↑	Insight↑	Stnd.↑	Util.↑	Overall↑	Rate↑	Effect.↑	Fact.↑	Overall↑
LLM with Search Tools										
Claude-Sonnet 4	0.427	0.664	0.609	0.669	0.614	0.641	0.342	0.169	0.488	0.336
Claude-Sonnet 4 (Think)	0.396	0.662	0.592	0.647	0.600	0.627	0.318	0.099	0.438	0.297
Gemini-3-Flash	0.484	0.700	0.650	0.712	0.639	0.678	0.319	0.354	0.669	0.401
Gemini-3-Flash (Think)	0.464	0.683	0.629	0.705	0.627	0.664	0.305	0.339	0.611	0.378
GPT-4.1	0.445	0.666	0.625	0.682	0.632	0.652	0.301	0.288	0.570	0.356
GPT-5	0.472	0.711	0.689	0.737	0.648	0.700	0.340	0.296	0.544	0.374
GPT-5.2	0.527	0.745	0.728	0.772	0.682	0.736	0.353	0.405	0.693	0.437
grok-4	0.486	0.673	0.632	0.684	0.614	0.654	0.296	0.383	0.760	0.414
Baichuan-M2-Plus	0.529	0.710	0.656	0.697	0.654	0.681	0.338	0.311	0.951	0.463
Baichuan-M3-Plus	0.498	0.685	0.636	0.683	0.632	0.661	0.283	0.269	**0.971**	0.427
Deep Research Agents										
Tongyi DR-30B	0.348	0.563	0.542	0.533	0.573	0.551	0.189	0.236	0.475	0.261
Kimi-K2.5-Agent	0.540	0.730	0.668	0.714	0.653	0.695	0.415	0.415	0.689	0.474
AgentScope	0.553	0.732	0.692	0.736	0.688	0.714	0.405	0.395	0.781	0.484
O4-mini DR	0.561	0.726	0.673	0.738	0.641	0.699	0.444	0.254	0.908	0.502
Perplexity Sonar DR	0.576	0.699	0.650	0.704	0.615	0.672	0.447	0.450	0.857	0.536
MiroThinker-v1.5	0.614	0.780	0.743	0.779	0.695	0.754	0.433	0.483	0.946	0.554
MiroThinker-v1.5-pro	**0.631**	**0.803**	**0.757**	**0.800**	**0.704**	**0.772**	**0.452**	**0.496**	0.960	**0.570**

Table 1: Overall performance on MedProbeBench. *Holistic Quality*: Compl. (Comprehensiveness), Insight (Insight Depth), Stnd. (Standards Adherence), Util. (Utility). *Fine-grained Verification*: Rate (Task Success Rate), Effect. (Search Effectiveness), Fact. (Factual Consistency).

Agents, including o4-mini-deep-research (OpenAI, 2025a), Perplexity Sonar Deep Research (Perplexity AI, 2025), and Kimi-K2.5-Agent (Moonshot AI, 2024); and (3) *Open-source Deep Research Agents*, including AgentScope-GPT4.1 (Gao et al., 2024), MiroThinker-v1.5, MiroThinker-v1.5-pro (Team et al., 2025a), and Tongyi-DeepResearch-30B-A3B (Team et al., 2025b).

Evaluation Protocol. MedProbeBench contains 50 tasks, each corresponding to one tumor entity from five WHO Classification volumes (Appendix E). For each task, a system receives the disease name, required section schema, and output constraints, then generates one complete guideline under zero-shot structured prompting. We evaluate each output with MedProbe-Eval, using GPT-4.1 (OpenAI, 2025c) as the automated judge with fixed prompts and zero temperature, and retrieve cited reference content through the JINA Reader API. Generation and evaluation prompts are provided in Appendix B. Because deep research evaluation is costly, each system is evaluated once per task; reliability is supported by aggregation over 50 tasks and more than 5,130 claim-level judgments, with additional validation reported in Appendix C.

5.2 Findings on Overall Performance

Conclusion 1: Deep Research Agents demonstrate architectural advantages over LLM+Search paradigms. Table 1 shows a clear advantage for strong deep research agents: MiroThinker-v1.5, MiroThinker-v1.5-pro, and Perplexity Sonar DR occupy the top three positions, indicating the value of iterative retrieval and evidence aggregation for guideline generation. This advantage is not uniform, however. Baichuan-M2-Plus is the strongest LLM+Search baseline and approaches DR-agent performance, while Tongyi DR-30B underperforms several LLM+Search systems, suggesting that system design quality matters as much as the broad agent paradigm.

Insufficiency 1: Fine-grained evidence verification remains the primary bottleneck across all model architectures. All systems score substantially lower on *Fine-grained Evidence Verification* than on *Holistic Rubrics*, exposing a gap between producing plausible guideline-style text and grounding claims in appropriate evidence. Low *Task Success Rate* indicates incomplete coverage of expert-identified claims, while weak *Search Effectiveness* suggests difficulty retrieving guideline-relevant sources rather than loosely related references. *Factual Consistency* is stronger for some DR agents once relevant evidence is available, but extended thinking alone does not reliably improve grounding, underscoring the need for retrieval and citation-integration mechanisms rather than reasoning depth alone.

5.3 Findings on Knowledge Types

Conclusion 2: Systems perform best on factual and clinical claims but struggle with specialized

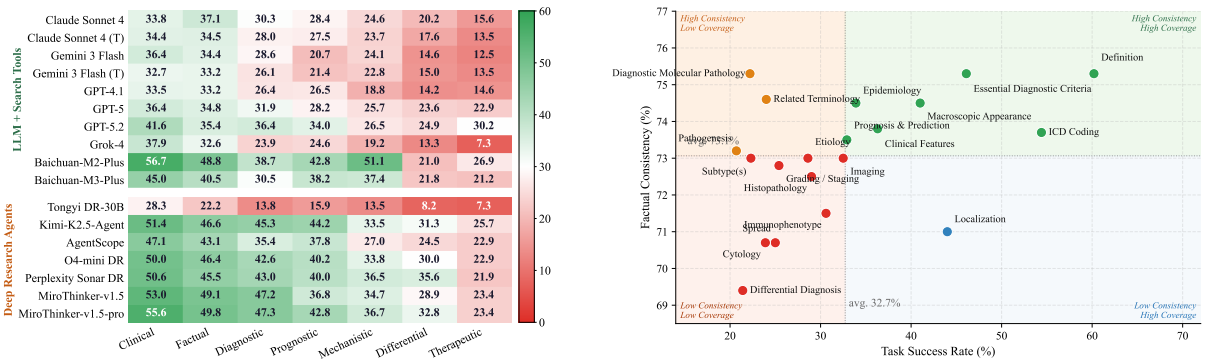

Figure 5: **Left:** Task Success Rate (%) by knowledge type across all evaluated systems. Darker blue indicates higher coverage. **Right:** Coverage vs. Consistency by guideline section type. Citation consistency remains uniformly high while claim coverage varies dramatically across sections.

reasoning. Figure 5 (left) shows substantial variation across knowledge types. Systems perform best on factual and clinical claims, suggesting relative strength in descriptive medical knowledge and routine clinical documentation. Diagnostic and prognostic claims are more difficult, but leading systems still show meaningful gains: Baichuan-M2-Plus reaches 56.7% on clinical claims, while MiroThinker-v1.5-pro reaches 47.3% on diagnostic claims.

Insufficiency 2: Critical performance gaps in mechanistic reasoning and treatment guidance severely limit clinical utility. Performance drops sharply for reasoning-intensive knowledge types: average accuracy is 28.8% for mechanistic claims, 22.2% for differential claims, and 19.2% for therapeutic claims. These categories require synthesizing molecular mechanisms, genotype–phenotype relationships, comparative diagnosis, and treatment evidence. Although Baichuan-M2-Plus reaches 51.1% on mechanistic claims, therapeutic accuracy remains below 30% even for the best systems, indicating that evidence-based treatment synthesis remains a major unresolved weakness.

5.4 Findings on Section Types

Conclusion 3: Citation consistency does not guarantee adequate evidence grounding. Figure 5 (right) shows that citation consistency and claim coverage diverge across section types. Many sections maintain high citation–claim consistency among generated claims, yet coverage drops in diagnosis- and mechanism-oriented sections. This indicates that systems can preserve citation formatting and local consistency while still omitting key expert claims.

Insufficiency 3: Diagnostic reasoning failures undermine reliability for clinical decision support. The weakest sections are those requiring comparative diagnostic reasoning rather than factual enumeration. Systems can often recover standard diagnostic criteria, but they struggle with differential diagnosis and mechanistic interpretation, where correct exclusion and comparison are central. This failure mode is clinically important because a guideline may appear complete while still missing the distinctions needed for reliable decision support.

Representative case analysis. A case-level walkthrough further illustrates why document-scale evaluation is necessary. In the ASPLT example (Appendix G), GPT-4.1 generates more claims than the gold guideline (76 vs. 56) and includes a similar proportion of cited claims (43.4% vs. 42.9%). However, its distribution collapses toward factual statements (71/76, 93.4%), while diagnostic, clinical, and mechanistic claim types present in the gold guideline are entirely missing. This example shows that longer outputs and frequent citations can still fail to recover the expert reasoning structure required by clinical guidelines.

5.5 Data Contamination Risk Analysis

Because the WHO Classification of Tumours volumes serve as the authoritative reference corpus underlying our benchmark, models exposed to these texts during pre-training could in principle inflate performance through verbatim memorisation. We quantify this risk via a *prefix-completion probing* test inspired by membership-inference literature (Carlini et al., 2021): each model is given the first half of a curated WHO passage and asked to continue verbatim, with no retrieval or system context. We sample 150 passages (30 per volume,

80–150 words, stratified over diagnostic criteria, molecular pathology, immunohistochemistry, and prognosis), score completions against the withheld continuations with ROUGE-L and BLEU-4 at temperature 0, and compare against a word-shuffled baseline that bounds chance similarity.

Model	ROUGE-L↓	BLEU-4↓	Risk
Random-shuffle	0.072 ± 0.019	0.018 ± 0.008	–
Claude-Sonnet 4	0.131 ± 0.041	0.042 ± 0.017	Safe
Gemini-3-Flash	0.158 ± 0.048	0.051 ± 0.021	Safe
GPT-4.1	0.174 ± 0.053	0.058 ± 0.022	Safe
GPT-5	0.196 ± 0.057	0.067 ± 0.025	Safe
GPT-5.2	0.213 ± 0.060	0.074 ± 0.027	Safe
Baichuan-M3-Plus	0.147 ± 0.044	0.046 ± 0.019	Safe
MiroThinker-v1.5-pro	0.189 ± 0.055	0.063 ± 0.024	Safe

Table 2: Verbatim-completion scores for representative systems.

As shown in Table 2, every system falls in the safe zone: ROUGE-L stays within 0.131–0.213 and BLEU-4 within 0.042–0.074, both well below the 0.30 partial-contamination threshold. Although completions are marginally above the shuffled baseline, the absolute gap is consistent with general semantic coherence rather than verbatim recall. Three design choices further bound contamination impact: (i) the task requires *generating* structured guidelines rather than reproducing source text; (ii) claim-level evidence verification penalises outputs lacking externally retrievable citations, which memorisation alone cannot satisfy; and (iii) the 50 entities include uncommon and recently reclassified subtypes that occupy low-frequency niches in pretraining corpora. The system-level differences in Table 1 therefore reflect genuine deep-research capability rather than differential memorisation.

5.6 Reliability and Human-AI Alignment Validation

We validate MedProbe-Eval against expert judgment at both the document and claim levels. For **holistic quality**, three senior medical experts independently score 40 guideline samples (4 systems $\times$ 5 categories $\times$ 2 guidelines) using the same rubric and task-level weights as the automated judge; GPT-4.1 preserves the experts' relative ordering with statistically significant rank agreement (Spearman $\rho = 0.754$, $p = 0.012$, perfect directional agreement across all five dimensions), differing only in absolute calibration. For **claim-level grounding**, on 200 stratified claim-reference pairs the LLM verifier reaches F1 $= 86.4\%$ and Cohen's $\kappa = 0.79$ against majority-vote expert labels. A multi-judge cross-check with GPT-4.1, Claude-4-Sonnet, and Gemini-2.5-Pro yields ICC $= 0.944$ (holistic) and 0.931 (evidence verification); residual same-family calibration drift is small and does not alter system-level rankings. Full study design, statistics, and per-judge results are reported in Appendix D and Appendix F (see Table 3, Table 4, and Figure 6).

6 Conclusion

We introduced **MedProbeBench**, a benchmark for evaluating deep research through the generation of evidence-grounded expert medical documentation, namely structured clinical content where every key claim must be supported by verifiable, externally retrievable citations. Our results show that although state-of-the-art models produce fluent and well-structured outputs, reliable evidence grounding remains a persistent bottleneck, especially for mechanistic reasoning and treatment-related content. High surface fluency and citation presence frequently mask substantive failures in specialized knowledge integration, revealing a critical gap between the appearance and the reality of evidence grounding. As medical AI systems move toward high-stakes clinical use, closing this synthesis-grounding gap is essential for trustworthy AI-assisted decision-making. MedProbeBench offers a standardized framework to expose these limitations and drive progress toward AI systems that meet the citation accountability standards of expert medical documentation.

Limitations

MedProbeBench focuses on evidence-grounded clinical guideline generation in oncology-related domains. Although it covers multiple disease systems and evaluates both holistic quality and fine-grained evidence verification, it does not exhaust all possible clinical specialties or guideline-writing scenarios. In addition, some reference-accessibility checks may be affected by publisher access restrictions, metadata inconsistencies, or URL resolution failures. Finally, while LLM-based judgments are used under structured rubrics and validation protocols, they may still introduce residual evaluation noise. MedProbeBench is intended only for research evaluation of evidence-grounded medical deep research systems, not for clinical advice and deployment certification.

References

Anthropic. 2025. Introducing Claude 4. https://www.anthropic.com/news/claude-4.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Baichuan Intelligence. 2025. Baichuan-M2-Plus. https://www.baichuan-ai.com/home.

Baichuan Intelligence. 2026. Baichuan-M3-Plus. https://www.baichuan-ai.com/home.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Shan Chen, Pedro Moreira, Yuxin Xiao, Sam Schmidgall, Jeremy Warner, Hugo Aerts, Thomas Hartvigsen, Jack Gallifant, and Danielle S Bitterman. 2025. Medbrowsecomp: Benchmarking medical deep research and computer use. *arXiv preprint arXiv:2505.14963*.

Google DeepMind. 2024. *Gemini Deep Research: Advanced AI research capabilities*. Https://blog.google/products/gemini/google-gemini-deep-research/.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*.

Dennis Fast, Lisa C Adams, Felix Busch, Conor Fallon, Marc Huppertz, Robert Siepmann, Philipp Prucker, Nadine Bayerl, Daniel Truhn, Marcus Makowski, and 1 others. 2024. Autonomous medical evaluation for guideline adherence of large language models. *NPJ Digital Medicine*, 7(1):358.

Nima Fathi, Amar Kumar, and Tal Arbel. 2025. Aura: A multi-modal medical agent for understanding, reasoning and annotation. pages 105–114.

Dawei Gao, Zitao Li, Xuchen Pan, Weirui Kuang, Zhijian Ma, Bingchen Qian, Fei Wei, Wenhao Zhang, Yuexiang Xie, Daoyuan Chen, and 1 others. 2024. Agentscope: A flexible yet robust multi-agent platform. *arXiv preprint arXiv:2402.14034*.

Shanghua Gao, Richard Zhu, Zhenglun Kong, Ayush Noori, Xiaorui Su, Curtis Ginder, Theodoros Tsiligkaridis, and Marinka Zitnik. 2025. Txagent: An ai agent for therapeutic reasoning across a universe of tools. *arXiv preprint arXiv:2503.10970*.

Eun Jeong Gong, Chang Seok Bang, Jae Jun Lee, and Gwang Ho Baik. 2025. Knowledge-practice performance gap in clinical large language models: systematic review of 39 benchmarks. *Journal of Medical Internet Research*, 27:e84120.

Google. 2025. Introducing Gemini 3 Flash: Benchmarks, global availability. https://blog.google/products/gemini/gemini-3-flash/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, and 1 others. 2025. Biomni: A general-purpose biomedical ai agent. *biorxiv*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Xiaomin Li, Mingye Gao, Yuexing Hao, Taoran Li, Guangya Wan, Zihan Wang, and Yijun Wang. 2025. Medguide: Benchmarking clinical decision-making in large language models. *arXiv preprint arXiv:2505.11613*.

XQ Li, W Cheuk, SS Chuang, and KL Rech. 2024. *WHO Classification of Tumours. Haematolymphoid Tumors*. IARC Press.

Jiyao Liu, Jinjie Wei, Wanying Qu, Chenglong Ma, Junzhi Ning, Yunheng Li, Ying Chen, Xinzhe Luo, Pengcheng Chen, Xin Gao, and 1 others. 2025. Medq-bench: Evaluating and exploring medical image quality assessment abilities in mllms. *arXiv preprint arXiv:2510.01691*.

David N Louis, Arie Perry, Pieter Wesseling, Daniel J Brat, Ian A Cree, Dominique Figarella-Branger, Cynthia Hawkins, HK Ng, Stefan M Pfister, Guido Reifenberger, and 1 others. 2021. The 2021 who classification of tumors of the central nervous system: a summary. *Neuro-oncology*, 23(8):1231–1251.

Moonshot AI. 2024. Kimi agent.

Iris D Nagtegaal, Robert D Odze, David Klimstra, Valerie Paradis, Massimo Rugge, Peter Schirmacher, Kay M Washington, Fatima Carneiro, Ian A Cree, and 1 others. 2019. The 2019 who classification of tumours of the digestive system. *Histopathology*, 76(2):182.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Junzhi Ning, Jiashi Lin, Yingying Fang, Wei Li, Jiyao Liu, Cheng Tang, Chenglong Ma, Wenhao Tang, Tianbin Li, Ziyan Huang, and 1 others. 2026. Mmrarebench: A rare-disease multimodal and multi-image medical benchmark. *arXiv preprint arXiv:2604.10755*.

WHO Classification of Tumours Editorial Board and 1 others. 2020. *Soft tissue and bone tumours*, volume 3. World Health Organization.

OpenAI. 2024. *Introducing Deep Research: AI agent for comprehensive research tasks*. Https://openai.com/index/introducing-deep-research/.

OpenAI. 2025a. Introducing deep research. https://openai.com/index/introducing-deep-research/.

OpenAI. 2025b. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/.

OpenAI. 2025c. *Introducing GPT-4.1 in the API*. Accessed: 2026-01-29.

OpenAI. 2025d. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/.

OpenAI. 2025e. Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/.

Perplexity AI. 2025. Introducing perplexity deep research: An ai-powered deep research assistant. Perplexity Deep Research accelerates in-depth question answering by performing iterative search, reasoning, and synthesis of multiple sources into a structured report.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, and 1 others. 2025. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. *arXiv preprint arXiv:2511.07685*.

Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Sibo Ju, Jin Ye, Pengcheng Chen, Ming Hu, and 1 others. 2025. Gmai-vl-r1: Harnessing reinforcement learning for multimodal medical reasoning. *arXiv preprint arXiv:2504.01886*.

MiroMind Team, Song Bai, Lidong Bing, Carson Chen, Guanzheng Chen, Yuntao Chen, Zhe Chen, Ziyi Chen, Jifeng Dai, Xuan Dong, and 1 others. 2025a. Mirothinker: Pushing the performance boundaries of open-source research agents via model, context, and interactive scaling. *arXiv preprint arXiv:2511.11793*.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025b. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.

Ming-Sound Tsao, Andrew G Nicholson, Joseph J Maleszewski, Alexander Marx, and William D Travis. 2022. *Reprint of "Introduction to 2021 WHO classification of thoracic tumors"*, volume 17. Elsevier.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, and 1 others. 2024. Mineru: An open-source solution for precise document content extraction.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2025a. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2025b. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.

xAI. 2025. Grok 4. https://x.ai/news/grok-4.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

Dingkang Yang, Jinjie Wei, Mingcheng Li, Jiyao Liu, Lihao Liu, Ming Hu, Junjun He, Yakun Ju, Wei Zhou, Yang Liu, and 1 others. 2025. Medaide: information fusion and anatomy of medical intents via llm-based agent collaboration. *Information Fusion*, page 103743.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

Appendix

A LLM-based Evaluation Protocol

A.1 Holistic Rubrics

This prompt is used to evaluate the overall quality of a generated medical guideline by directly comparing it against a gold-standard reference guideline under a strict GT-anchored scoring protocol. For each task, before scoring any system outputs, the same judge model is additionally prompted to produce task-specific weights over the four fixed holistic dimensions based only on the disease specification and required section schema, following a task-adaptive weighting strategy inspired by DeepResearch Bench (Du et al., 2025). The weight generation step is repeated for T trials and averaged, and the resulting task-level weights are then shared across all systems for that task.

```
You are a senior medical guideline reviewer conducting a STRICT
      BENCHMARK evaluation.

Your task is to score a GENERATED medical guideline by DIRECTLY
      COMPARING it to a REFERENCE guideline (ground truth).

## Reference Guideline (Ground Truth, Score = 10):
{reference_content}

## Generated Guideline (To Be Scored by Deviation):
{generated_content}

## Evaluation Criteria (Each scored independently, GT = 10):
{criteria_list}

IMPORTANT ANCHOR RULE:
- The REFERENCE guideline represents perfect, expert-level quality and
      is implicitly scored as 10/10.
- The GENERATED guideline is scored ONLY by how much it DEVIATES from
      the reference.
- Scores MUST be assigned by starting from 10 and subtracting points
      based on differences, weaknesses, or omissions.

## Relative Scoring Philosophy (GT-ANCHORED):
- 10: Generated guideline is effectively identical to the reference.
- 8-9: Minor deviations from the reference.
- 6-7: Noticeable deviations.
- 4-5: Major deviations.
- 2-3: Severe deviations.
- 0-1: Extreme deviation.

## Mandatory Benchmark Rules:
1. GT = 10 by definition.
2. Score ONLY the distance between GENERATED and REFERENCE.
3. Any missing, weaker, or less specific content MUST reduce the score.
4. Vagueness counts as deviation when the reference is specific.
5. Equal section titles do NOT imply equivalence.
6. When uncertain, subtract points.

## Output Requirements:
- Output ONLY a valid JSON object.
- No explanations, no markdown, no additional text.
- Use EXACTLY the following structure:

{"scores": [
  {"criterion": "<criterion name>", "score": <integer 0-10>, "reason":
      "<difference from GT>"}
]}
```

A.2 Fine-grained Evidence Verification

This section details the full fine-grained evidence verification pipeline. We first describe the claim matching threshold used in semantic matching (§A.2.2), then the claim extraction prompt that decomposes guidelines into atomic claims (§A.2.3). We then present three independent prompt templates for evaluating evidence grounding at the claim and reference level, corresponding to task success rate (§A.2.4), search effectiveness (§A.2.5), and factual consistency (§A.2.6).

A.2.1 Formal Claim Representation

To support claim-level evaluation, each guideline $\mathcal{G}$ is decomposed into a set of atomic claims by an extraction operator $\mathcal{D} : \mathcal{G} \rightarrow \mathcal{C}$, where $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$. Each claim is represented as $c_i = (t_i, \mathcal{R}_i, \tau_i)$, where t_i is the claim text, $\mathcal{R}_i$ is the set of cited references associated with the claim, and τ_i is the claim type. Claim types are drawn from seven clinically meaningful categories: Factual, Mechanistic, Clinical, Diagnostic, Differential, Prognostic, and Therapeutic.

For matching generated and gold claims, we define a semantic relation $\sim$: $c_1 \sim c_2$ if $\mathrm{sim}(t_1, t_2) \geq \theta$, where $\mathrm{sim}(\cdot, \cdot)$ denotes the semantic similarity between claim texts. A generated claim is counted as matched if it is semantically equivalent to at least one gold claim. The matched generated set is

$$\mathcal{C}^{\mathrm{hit}} = \left\{ c \in \mathcal{C}^{\mathrm{gen}} \,\middle|\, c \sim \mathcal{C}^{\mathrm{gold}} \right\}. \tag{6}$$

Before computing this set, semantically duplicate generated claims are merged so that each gold claim can be credited at most once.

A.2.2 Claim Matching Threshold

In the two-stage claim matching procedure, we use a Jaccard similarity threshold $\theta = 0.85$ as a conservative pre-filter. Claims with a Jaccard score at or above θ are considered matched directly; all remaining pairs are passed to an LLM-based semantic evaluator. This threshold is intentionally set high so that only claims with exceptionally high lexical overlap bypass LLM evaluation, ensuring that all borderline cases receive semantic judgment. Since the actual matching decision for most claim pairs is made by the LLM, the specific value of θ has minimal impact on final evaluation quality.

A.2.3 Claim Extraction

This prompt is applied to each normalized section of a guideline (both gold-standard and generated) to extract atomic, independently verifiable clinical claims. Each claim is returned as a structured record containing the claim text, associated citation markers, and a knowledge-type label.

```
# Medical Claim Extraction

## Task
Extract medical claims from the following medical guideline section.
```

```
Each claim must be a minimal, independently verifiable factual
       statement.

## Section Text
{section_text}

## Output Format
Return a JSON object with the following structure:
{
  "claims": [
    {
      "id": "C001",
      "content": "Extracted factual statement",
      "reference": "[1]",
      "type_knowledge": "Factual",
      "section": "<section name>"
    }
  ]
}

## Field Definitions
- id: Sequential identifier (C001, C002, ...)
- content: Claim text (preserve original wording, values, and
       annotations)
- reference: In-text citation markers such as "[1]", "[2, 3]", or "" if
        absent
- type_knowledge: One of [Factual, Mechanistic, Clinical, Diagnostic,
                          Differential, Prognostic, Therapeutic]
- section: The section name from which the claim was extracted

## Knowledge Type Guide
- Factual: Descriptive facts, statistics, anatomical locations, ICD
       codes
- Mechanistic: Causal mechanisms, molecular pathways, genetic
       alterations
- Clinical: Clinical manifestations, signs, symptoms
- Diagnostic: Diagnostic tests, imaging or pathology findings, criteria
- Differential: Features distinguishing a disease from related
       conditions
- Prognostic: Survival statistics, outcome predictors, recurrence rates
- Therapeutic: Treatment methods, drug regimens, efficacy or toxicity
       data

## Extraction Rules
1. One fact per claim: split compound sentences into atomic units
2. Preserve details: retain all quantitative values, percentages, and
       ranges
3. Preserve aliases: e.g., "SMARCB1 (also known as hSNF5, INI1, or
       BAF47)"
4. Omit figure and table cross-references (e.g., "see Figure 1")
5. Do not paraphrase: keep the original medical terminology
```

A.2.4 Task Success Rate

Task Success Rate measures the proportion of gold-standard claims successfully covered by generated claims:

$$S_1(\hat{\mathcal{G}}) = \frac{|\mathcal{C}^{\text{hit}}|}{|\mathcal{C}^{\text{gold}}|}. \tag{7}$$

The following prompt determines whether a gold-standard (GT) claim is semantically covered by any predicted claim.

```
Determine if the GT claim is semantically covered by any predicted
       claim.

GT Claim: {gt_claim}

Predicted Claims:
{pred_claims_json}

Match if: same medical fact, equivalent meaning, or prediction covers
       GT's core info.

Return JSON only:
{"is_hit": true/false, "matched_pred_claim_id": "id or empty"}
```

A.2.5 Search Effectiveness

Search Effectiveness evaluates whether generated references recover the evidence used by the gold guideline. For each section s, let $\mathcal{R}_s^{\text{gold}}$ and $\mathcal{R}_s^{\text{gen}}$ denote the gold and generated reference sets. References are matched by content-based semantic comparison, with each generated reference matched to at most one gold reference. Let M_s denote the matched gold references in section s; reference recall is

$$\text{Recall} = \frac{\sum_s |M_s|}{\sum_s |\mathcal{R}_s^{\text{gold}}|}. \tag{8}$$

To account for the amount of retrieved evidence, we define

$$\text{Quantity} = \min\left(1, \frac{\sum_s |\mathcal{R}_s^{\text{gen}}|}{\sum_s |\mathcal{R}_s^{\text{gold}}|}\right), \tag{9}$$

and compute $S_2(\hat{\mathcal{G}}) = 0.6 \cdot \text{Recall} + 0.4 \cdot \text{Quantity}$. The following prompt evaluates whether a predicted reference corresponds to the same underlying object as the gold-standard reference.

```
Determine whether the following two URLs refer to the same underlying
       topic, work, or object.

GT URL: {gt_url}
GT Content (excerpt):
{gt_content}

Pred URL: {pred_url}
Pred Content (excerpt):
{pred_content}

Consider it a match if both contents discuss the same underlying thing
       (e.g., the same study, dataset, method, system, or phenomenon),
       even if their wording, focus, conclusions, or level of detail
       differ.

If one or both contents are mainly meta-information (e.g., titles,
       abstracts, or repository pages), use available clues (such as
       names or referenced objects) and apply a slightly more
       permissive judgment.

Ignore minor differences or disagreements.

Return JSON only:
{"is_match": true/false, "reason": "brief explanation of the shared
       underlying object"}
```

A.2.6 Factual Consistency

Factual Consistency evaluates whether generated claims are grounded in accessible cited content. For each generated claim c, we resolve its references into URLs $\mathcal{R}^{\text{url}}(c)$, retrieve their contents, and retain the accessible subset $\mathcal{R}^{\text{acc}}(c)$. Let $\mathcal{C}^{\text{url}} = \{c \in \mathcal{C}^{\text{gen}} \mid |\mathcal{R}^{\text{url}}(c)| > 0\}$ be the claims with at least one resolvable URL. We define

$$\text{verify}(c) = \begin{cases} 1, & |\mathcal{R}^{\text{acc}}(c)| > 0 \\ & \wedge\ \text{rel}(c, \mathcal{R}^{\text{acc}}(c)) = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

where $\text{rel}(\cdot)$ is an LLM-based judgment of whether the cited content supports the claim. The factual consistency score is

$$S_3(\hat{\mathcal{G}}) = \frac{1}{|\mathcal{C}^{\text{url}}|} \sum_{c \in \mathcal{C}^{\text{url}}} \text{verify}(c). \tag{11}$$

The following prompt assesses whether a generated claim is actually supported by its cited reference

content, that is, whether the specific factual assertion made in the claim can be substantiated by the evidence in the reference, rather than merely sharing a topical area.

```
Determine whether the cited reference content provides evidence that
      supports the specific claim.

Claim:
{claim}

Reference Content:
{reference_content}

Your task is to judge evidential support, not topical overlap.

Return TRUE only if the reference content contains information that
      substantiates the specific factual assertion in the claim.
This includes: numerical values, diagnostic criteria, molecular
      findings, prevalence data, prognosis estimates, or mechanistic
      relationships that are stated or clearly implied in the
      reference.

Return TRUE if the reference is only an abstract or brief metadata, but
       the abstract explicitly reports findings consistent with the
      claim (e.g., the abstract reports the same gene mutation, the
      same survival statistic, or the same diagnostic criterion stated
       in the claim).

Return FALSE if:
- The claim makes a specific factual assertion (e.g., a percentage, a
      molecular marker, a treatment outcome) that is not mentioned or
      supported in the reference content.
- The reference only covers the same broad disease area or topic
      without providing evidence for the specific claim.
- The reference contradicts or is inconsistent with the claim.
- The reference is purely meta-information (title/landing page only)
      with no substantive findings that match the claim.

Return JSON only:
{"is_consistent": true/false}
```

B Guideline Generation Prompt Template

Below we present the prompt template used for medical guideline generation tasks. The template enforces structured output, evidence usage constraints, and explicit citation requirements.

```
# System Role
{role_and_constraints}

# Task
Write a comprehensive medical guideline chapter for: **{tumor_name}**

# Evidence & Literature Guidance
- You are encouraged to leverage scientific literature (e.g., PubMed-
      indexed studies) to enhance the accuracy, depth, and credibility
       of the content.
- When relevant, distinguish evidence derived from:
  - Randomized controlled trials (RCTs)
  - Cohort or case-control studies
  - Exploratory or molecular/mechanistic studies
- Clearly reflect the strength and limitations of available evidence.

# Literature Usage Constraints
- Only cite articles with verifiable PubMed IDs or DOIs. Do not
      fabricate references.
- ONLY search for original research papers published in peer-reviewed
      journals.

# Citation Rules
- Every factual, clinical, or scientific claim must be supported by at
      least one appropriate citation, unless explicitly stated as
      expert opinion or lacking available evidence.
- Use inline numeric citations in square brackets, e.g., [1], [2], or
      [1, 2, 3].

# Required Sections
{sections_text}

# Output Format
Output in Markdown format:

# {tumor_name}

## Section Name
Structured, guideline-style content with clear clinical relevance.
Evidence-based discussion with inline citations.
Key statistics, outcomes, and limitations where applicable...

...(Continue for all sections - remember citations in EVERY section)

## References
1. Author et al. Title. *Journal*. Year;Volume:Pages. URL: https://
      pubmed.ncbi.nlm.nih.gov/XXXXXXXX/
2. Author et al. Title. *Journal*. Year;Volume:Pages. URL: https://doi.
      org/10.XXXX/XXXXX

...(List all cited references)
```

C Evaluation Reliability

Due to the high computational cost of deep research evaluation, each system is evaluated once per task. Result reliability is supported by three factors: (1) each system's score aggregates over approximately 5,130 claim-level judgments across 50 tasks, providing fine-grained statistical coverage beyond task-level counts; (2) performance rankings remain consistent across all five medical domains (see Appendix E), indicating stable cross-domain conclusions; and (3) single-round evaluation is standard practice in comparable long-form research benchmarks (Du et al., 2025).

D Human–AI Alignment Analysis

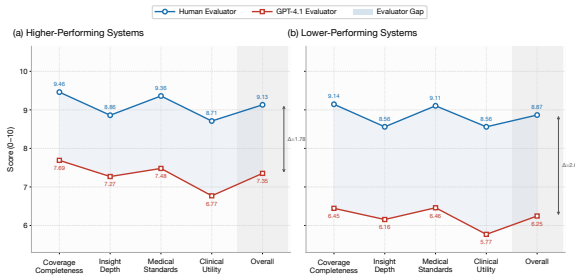

Figure 6: Human and GPT-4.1 evaluation scores for higher-performing Good systems and lower-performing Bad systems across four dimensions and the overall score.

This section reports the detailed quantitative results underlying the human–AI alignment validation presented in Section 5. We validate both tiers of MedProbeBench's evaluation framework: holistic rubric scoring (§D.1) and fine-grained evidence verification (§D.2). In each case, we summarize the study design, report agreement statistics, and discuss implications for the reliability of automated evaluation.

D.1 Holistic Rubric Alignment

Following the evaluation protocol described in Section 5, three senior medical experts independently evaluate 40 guideline samples (4 systems $\times$ 5 categories $\times$ 2 guidelines) using the same four-dimensional rubric and task-level dynamic weights

Metric	Value
Spearman ρ	0.754
p-value	0.012
Directional agreement	5/5 dimensions
Mean absolute difference	2.21

Table 3: Human–AI alignment statistics for holistic rubric evaluation. Scores are on a 0–10 scale.

Metric	Value
Precision	87.3%
Recall	85.6%
F1	86.4%
Cohen's κ	0.79

Table 4: Human validation of automated factual consistency verification against majority-vote expert labels.

as the automated GPT-4.1 judge. The four dimensions are Medical Coverage Completeness, Mechanistic Insight Depth, Adherence to Medical Standards, and Clinical Utility; each dimension is assessed through fine-grained criteria rather than a single coarse score. Human experts score each criterion on a 0–10 scale, and scores are aggregated using the same task-level weights as automated evaluation. These weights are generated once from the task specification and section schema, then fixed across all systems and all human evaluations for that task. The sampled systems cover both stronger and weaker performers in Table 1: MiroThinker-v1.5-pro and O4-mini DR are grouped as higher-performing systems, while Tongyi DR-30B and GPT-4.1 are grouped as lower-performing systems. To quantify agreement, we compute Spearman rank correlation between GPT-4.1 scores and aggregated human expert scores across all dimension–group combinations. Table 3 summarizes the results.

Both evaluators consistently rank higher-quality systems above lower-quality ones across all five evaluation dimensions, yielding perfect directional agreement. The mean absolute scoring difference of 2.21 points reflects a systematic calibration offset rather than disagreement in discriminative judgment: human experts tend to assign higher absolute scores, but the relative ranking of systems is preserved. These results collectively validate GPT-4.1 as a reliable proxy for expert judgment in holistic quality assessment.

D.2 Fine-grained Evidence Verification Alignment

As described in Section 5, we conduct a targeted human validation study to assess the reliability of LLM-based factual consistency verification. We randomly sample 200 claim-reference pairs from the full evaluation set, stratified across section types and model outputs to ensure broad coverage. Three senior medical experts independently annotate each pair as *supported* or *not supported*, based on whether the cited reference genuinely substantiates the associated clinical claim; majority vote serves as the gold standard. Results are summarized in Table 4.

These results confirm that GPT-4.1-based factual consistency verification reliably aligns with expert judgment, validating its use as an automated proxy for claim-level evidence grounding.

E Dataset Statistics

This section provides detailed statistics on the MedProbeBench dataset and supplementary evaluation breakdowns referenced from the main text. A high-level overview of guideline and claim counts per medical domain is shown in Figure 4 of the main text. We first describe dataset construction details (§E.1), followed by the ground-truth data composition: claim type distribution (§E.2), dataset composition by medical domain (§E.3), and benchmark scale statistics (§E.4). We then report fine-grained evaluation breakdowns that supplement the findings in Section 5: task success rate by knowledge type (§E.5) and system performance by section type (§E.6).

E.1 Dataset Construction Details

MedProbeBench is constructed from the WHO Classification of Tumours (5th Edition), using five oncology volumes that cover major organ systems: Central Nervous System Tumours, Soft Tissue and Bone Tumours, Digestive System Tumours, Thoracic Tumours, and Haematolymphoid Tumours. Source documents in PDF or HTML format are converted into structured Markdown using MinerU, then reviewed by board-certified medical professionals to ensure clinical reliability before downstream normalization.

To enable fair comparison between gold-standard guidelines and model-generated reports, both document types are processed through the same structuring pipeline: (1) format compliance

Knowledge Type	Count	Percentage
Clinical (Clin.)	379	7.38%
Factual (Fact.)	2422	47.18%
Diagnostic (Diag.)	1060	20.65%
Prognostic (Prog.)	443	8.63%
Mechanistic (Mech.)	482	9.39%
Differential (Diff.)	247	4.81%
Therapeutic (Ther.)	100	1.95%
Total	**5133**	**100%**

Table 5: Distribution of knowledge types in MedProbeBench.

checking against the required section schema; (2) Markdown structure normalization with standardized headings and hierarchy; (3) section-level segmentation; (4) atomic clinical claim extraction with associated in-text citation markers; and (5) citation normalization and enrichment, which resolves non-canonical citations into canonical PubMed- or DOI-based references when possible.

Across all tasks, the benchmark covers 20 guideline section types: Definition, ICD Coding, Essential Diagnostic Criteria, Localization, Macroscopic Appearance, Clinical Features, Grading / Staging, Spread, Diagnostic Molecular Pathology, Pathogenesis, Epidemiology, Prognosis and Prediction, Imaging, Immunophenotype, Histopathology, Etiology, Cytology, Subtype(s), Differential Diagnosis, and Related Terminology. These sections are further mapped into seven knowledge types used in claim-level analysis: Factual, Mechanistic, Clinical, Diagnostic, Differential, Prognostic, and Therapeutic.

E.2 Claim Type Distribution

We decompose all gold-standard guidelines into atomic claims following the methodology described in Section 4. Each claim is categorized into one of seven knowledge types according to its semantic content and clinical role. Table 5 summarizes the distribution of knowledge types across the entire benchmark.

E.3 Dataset Composition by Medical Domain

The benchmark encompasses five WHO classification volumes, covering diverse organ systems and disease categories. Table 6 summarizes the distribution of guidelines and claims across different medical domains.

Medical Domain	Guidelines	Claims	Avg. Claims/Guideline
Central Nervous System Tumours	9	970	107.78
Haematolymphoid Tumours	12	1033	86.08
Digestive System Tumours	9	1213	134.78
Thoracic Tumours	10	1181	118.10
Soft Tissue and Bone Tumours	10	736	73.60
Total	**50**	**5133**	**102.66**

Table 6: Dataset composition by medical domain.

E.4 Benchmark Scale Statistics

Table 7 reports the full scale of MedProbeBench's evaluation infrastructure: the number of task-adaptive holistic rubric criteria generated across all 50 tasks, broken down by dimension, alongside the citation coverage of the benchmark's ground-truth claims.

Task-Adaptive Weight Derivation. For holistic rubric evaluation, we derive task-specific dimension weights w_i from the task specification and required section schema, while keeping the four top-level medical dimensions fixed. For each task t, the judge LLM independently produces dimension weights $w_i^{(m,t)}$ over T trials, and we use the averaged weights:

$$w_i^{(t)} = \frac{1}{T} \sum_{m=1}^{T} w_i^{(m,t)}, \quad i \in \{1, 2, 3, 4\}, \quad (12)$$

followed by normalization so that $\sum_i w_i^{(t)} = 1$. The resulting weight vector is generated once per task and then shared by all evaluated systems on that task, ensuring consistent cross-system comparison.

Analogously, for evidence verification, we assign task-specific weights α_j to the three metrics based on the same task description and section schema. For each task t, these weights are averaged across T independent trials:

$$\alpha_j^{(t)} = \frac{1}{T} \sum_{m=1}^{T} \alpha_j^{(m,t)}, \quad j \in \{1, 2, 3\}, \quad (13)$$

and normalized to satisfy $\sum_j \alpha_j^{(t)} = 1$. As with $w_i^{(t)}$, the final $\alpha_j^{(t)}$ are fixed at the task level and reused for all systems evaluated on that task.

The 1,236 total rubric criteria underscore the fine-grained, disease-specific nature of MedProbeBench's holistic evaluation: each task receives its own rubric tailored to the disease's clinical priorities, rather than a single fixed template applied uniformly. The 42.92% citation rate in GT claims reflects the evidence grounding standard of

Component	Count	Details
Holistic Rubric Criteria (across 50 tasks)		
Comprehensiveness	352	
Insight Depth	296	
Accuracy & Standards	294	
Clinical Utility	294	
Total Rubric Criteria	**1,236**	≈24.7 per task
Ground-Truth Claim Statistics		
Total atomic claims	5,133	across 50 tasks
Claims with ≥1 citation	2,203	42.92% of total
Claims without citation	2,930	57.08% of total

Table 7: Benchmark scale statistics for MedProbeBench. Holistic rubric criteria are generated per-task and summed across all 50 tasks. Claims with references are GT claims carrying at least one supporting citation.

WHO guidelines: highly descriptive or definitional claims (terminology, ICD codes, macroscopic features) typically carry no citation, while mechanistic, prognostic, and diagnostic claims are expected to be evidence-backed.

E.5 Task Success Rate by Knowledge Type

Table 8 reports the full task success rate breakdown by knowledge type for all evaluated systems, supplementing the analysis in Section 5 (Findings on Knowledge Types).

Model	Clin.	Fact.	Diag.	Prog.	Mech.	Diff.	Ther.
LLM with Search Tools							
Claude Sonnet 4	33.8	37.1	30.3	28.4	24.6	20.2	15.6
Claude Sonnet 4 (T)	34.4	34.5	28.0	27.5	23.7	17.6	13.5
Gemini 3 Flash	36.4	34.4	28.6	20.7	24.1	14.6	12.5
Gemini 3 Flash (T)	32.7	33.2	26.1	21.4	22.8	15.0	13.5
GPT-4.1	33.5	33.2	26.4	26.5	18.8	14.2	14.6
GPT-5	36.4	34.8	31.9	28.2	25.7	23.6	22.9
GPT-5.2	41.6	35.4	36.4	34.0	26.5	24.9	**30.2**
grok-4	37.9	32.6	23.9	24.6	19.2	13.3	7.3
Baichuan-M2-Plus	**56.7**	48.8	38.7	42.8	**51.1**	21.0	26.9
Baichuan-M3-Plus	45.0	40.5	30.5	38.2	37.4	21.8	21.2
Deep Research Agents							
AgentScope	47.1	43.1	35.4	37.8	27.0	24.5	22.9
Tongyi DR-30B	28.3	22.2	13.8	15.9	13.5	8.2	7.3
O4-mini DR	50.0	46.4	42.6	40.2	33.8	30.0	22.9
Perplexity Sonar DR	50.6	45.5	43.0	40.0	36.5	**35.6**	21.9
kimi-agent	51.4	46.6	45.3	**44.2**	33.5	31.3	25.7
mirothinker-v1.5	53.0	49.1	47.2	36.8	34.7	28.9	23.4
mirothinker-v1.5-pro	55.6	**49.8**	**47.3**	42.8	36.7	32.8	23.4
Type Average	42.6	39.2	33.8	32.4	28.8	22.2	19.2

Table 8: Task Success Rate by knowledge type (%).

E.6 System Performance by Section Type

Table 9 reports task success rate and factual consistency across all 20 guideline section types, averaged over all systems, supplementing the analysis in Section 5 (Findings on Section Types).

F Multi-Judge Validation

To examine potential self-preference bias and confirm the reliability of GPT-4.1 as the automated judge, we conduct a multi-judge validation study using three independent LLM judges: GPT-4.1, Claude-4-Sonnet, and Gemini-2.5-Pro. We randomly sample 200 guideline evaluations drawn from diverse models and disease topics across the benchmark, and each judge independently scores all samples on both Holistic Quality and Fine-grained Evidence Verification.

Tables 10 and 11 report per-model scores from each judge, together with Intraclass Correlation Coefficient (ICC) and Mean Absolute Deviation (MAD) between GPT-4.1 and the other two judges. Overall, ICC values of 0.944 (Holistic) and 0.931 (Evidence Verification) indicate strong but not perfect agreement, consistent with known calibration differences across LLM judge families (Zheng et al., 2023). MAD values range from 0.007 to 0.013, reflecting non-trivial though modest scoring differences at the individual evaluation level.

We observe residual in-family tendencies: GPT-4.1 assigns marginally higher holistic scores to GPT-5 and GPT-5.2 compared to Claude and Gemini judges (mean $\Delta = +0.011$, range $+0.008$–$+0.015$), and Claude-4-Sonnet exhibits an analogous pattern for Claude-Sonnet 4 ($\Delta = +0.009$). These differences fall within the overall inter-judge MAD range and do not change the rank ordering of any system pair separated by more than 0.02 in the overall score. We therefore recommend treating score gaps smaller than 0.02 with caution, while noting that the principal findings and system-level conclusions reported in Section 5 remain robust across all three judges.

G End-to-End Evaluation Walkthrough

To address reviewer concerns about the lack of concrete evaluation examples, we present a complete end-to-end walkthrough of the MedProbeBench evaluation pipeline using a representative case: *Atypical Spindle Cell/Pleomorphic Lipomatous Tumour* (ASPLT). This entity was selected for its compact yet structurally complete guideline (20 sections, 56 ground-truth claims, 15 references), making it suitable for detailed illustration without overwhelming length.

G.1 Case Overview

The model generates *more* claims than the GT (76 vs. 56), yet as the evaluation below demonstrates, increased volume does not automatically translate into higher coverage of expert-identified evidence.

Section Type	Task Success Rate↑	Consistency↑	Section Type	Task Success Rate↑	Consistency↑
Definition	**60.2%**	75.3%	Epidemiology	33.9%	74.5%
ICD Coding	54.4%	73.7%	Prognosis and prediction	32.9%	73.5%
Essential diagnostic criteria	46.1%	75.3%	Imaging	32.5%	73.0%
Localization	44.0%	71.0%	Immunophenotype	30.6%	71.5%
Macroscopic appearance	41.0%	74.5%	Histopathology	29.0%	72.5%
Clinical features	36.3%	73.8%	Etiology	28.6%	73.0%
Grading / Staging	25.4%	72.8%	Cytology	23.9%	70.7%
Spread	25.0%	70.7%	Subtype(s)	22.3%	73.0%
Diagnostic molecular pathology	22.2%	75.3%	Differential diagnosis	21.4%	69.4%
Pathogenesis	20.7%	73.2%	Related terminology	24.0%	74.6%
Section Average	32.7%	73.1%			

Table 9: System performance by guideline section type. Task success rate measures claim coverage accuracy; consistency measures citation–claim alignment.

Model	GPT-4.1	Claude	Gemini	MAD_{Claude}	MAD_{Gemini}
MiroThinker-v1.5-pro	0.772	0.763	0.779	0.009	0.007
MiroThinker-v1.5	0.754	0.746	0.761	0.008	0.007
AgentScope	0.714	0.722	0.709	0.008	0.005
GPT-5.2†	0.736	0.724	0.730	0.012	0.006
GPT-5†	0.700	0.689	0.695	0.011	0.005
Perplexity Sonar DR	0.672	0.665	0.678	0.007	0.006
Claude-Sonnet 4†	0.641	0.650	0.636	0.009	0.005
Tongyi DR-30B	0.551	0.543	0.558	0.008	0.007
ICC (95% CI)				0.944 [0.935, 0.953]	
Overall MAD				0.009	0.006

Table 10: Multi-judge validation on Holistic Quality. Scores are overall holistic quality composites (0–1 scale). MAD is computed between GPT-4.1 and each alternative judge across the 200 sampled evaluations. †: models where same-family judge tends to assign marginally higher scores (see text).

Model	GPT-4.1	Claude	Gemini	MAD_{Claude}	MAD_{Gemini}
MiroThinker-v1.5-pro	0.570	0.558	0.576	0.012	0.006
MiroThinker-v1.5	0.554	0.543	0.561	0.011	0.007
AgentScope	0.484	0.474	0.493	0.010	0.009
GPT-5.2†	0.437	0.425	0.431	0.012	0.006
GPT-5†	0.374	0.362	0.369	0.012	0.005
Perplexity Sonar DR	0.536	0.526	0.543	0.010	0.007
Claude-Sonnet 4†	0.336	0.345	0.330	0.009	0.006
Tongyi DR-30B	0.261	0.253	0.268	0.008	0.007
ICC (95% CI)				0.931 [0.919, 0.943]	
Overall MAD				0.011	0.007

Table 11: Multi-judge validation on Fine-grained Evidence Verification. Scores are overall evidence verification composites (0–1 scale). †: models where same-family judge tends to assign marginally higher scores (see text).

G.2 Ground-Truth Reference Guideline

The complete WHO-style ground-truth guideline for the ASPLT case is reproduced below across all 20 mandatory sections. Sections absent for this entity (e.g. Imaging, Spread) are explicitly marked "None," reflecting the authentic benchmark structure against which model outputs are evaluated.

Property	GT	GPT-4.1 Pred
Number of sections	20	20
Total claims extracted	56	76
Claims with references	24 (42.9%)	33 (43.4%)
Unique reference entries	15	13
Content length (chars)	10,105	9,016

Table 12: Comparison of ground-truth (GT) and GPT-4.1 prediction statistics for the ASPLT case.

GT Guideline: Atypical Spindle Cell / Pleomorphic Lipomatous Tumour

Definition. Atypical spindle cell / pleomorphic lipomatous tumour is a benign adipocytic neoplasm, characterized by ill-defined tumour margins and the presence of variable proportions of mild to moderately atypical spindle cells, adipocytes, lipoblasts, pleomorphic cells, multinucleated giant cells, and a myxoid or collagenous extracellular matrix. It has a low tendency for local recurrence if incompletely excised. Unlike conventional atypical lipomatous tumours, there is no risk for dedifferentiation.

ICD-O Coding / ICD-11 Coding. 3857/0 Atypical spindle cell/pleomorphic lipomatous tumour. 2E80 & XH4E98 Benign lipomatous neoplasm & Spindle cell lipoma.

Related Terminology. Acceptable: atypical spindle cell lipoma. Not recommended: spindle cell liposarcoma; fibrosarcoma-like lipomatous neoplasm.

Subtype(s). None.

Localization. Atypical spindle cell / pleomorphic lipomatous tumours arise in the subcutis slightly more frequently than in deep (subfascial) somatic soft tissues, and only occasionally in intracavitary or visceral locations. The anatomical distribution is wide, predominating in the limbs and limb girdles [1, 2, 4]. The most common locations are the hand and foot and the thigh, followed by the shoulder and buttock, forearm, knee, lower leg, and upper arm. Less common locations are the head and neck, genital area, trunk, and back [1, 2, 3]. Rare sites of involvement include the larynx, mediastinum, retroperitoneum, trachea, and appendix [1].

Clinical Features. The tumour manifests as a persistent or enlarging soft tissue mass, nodule, or swelling, sometimes with tenderness [1].

Imaging. None.

Spread. None.

Epidemiology. Atypical spindle cell / pleomorphic lipomatous tumour occurs predominantly in middle-aged adults, with a peak incidence in the sixth decade of life, but can affect patients of any age (cases described in patients aged 6–87 years [1, 2, 3]). The large majority of patients are >30 years old. There is a slight male predominance.

Etiology. Unknown.

Pathogenesis. Deletions or losses of 13q14, including RB1 and its flanking genes RCBTB2, DLEU1, and IM2B, have been identified in a substantial subset of cases [1, 2, 4, 3, 5, 6]. In addition, monosomy 7 has been reported in some cases [1, 7].

Macroscopic Appearance. Grossly, atypical spindle cell / pleomorphic lipomatous tumours are unencapsulated, show a nodular or multinodular growth pattern, and demonstrate ill-defined tumour margins. Tumour size is variable (range: 0.5–28 cm; median: 5–8.5 cm) [1, 2].

Histopathology. A wide range of microscopic appearances can be observed, even regionally within the same lesion, depending on the relative proportions of atypical spindle cells, adipocytes, lipoblasts, and pleomorphic (multinucleated) cells, as well as the variable amount of collagenous and/or myxoid extracellular matrix [1, 2]. The adipocytic component has a predominantly mature morphology, with variation in adipocytic size and shape. Patchy, often mild to moderate adipocytic atypia with chromatin coarsening, nuclear enlargement, and focal binucleation or multinucleation can be observed [8, 9]. Morphologically, the lipoblasts can vary from small and univacuolated or bivacuolated to larger and multivacuolated (pleomorphic). Bizarre, hyperchromatic, and sometimes pleomorphic multinucleated cells are often scattered within the spindle cell or adipocytic components. Mitotic figures are often present but mostly scarce [2, 10, 11]. Tumour necrosis is absent.
The morphology of these tumours can best be described as a broad spectrum defined by two morphological extremes [1]. At the low-cellularity extreme, tumours can be paucicellular, with few, cytologically bland spindle cells with minimal nuclear atypia set in a prominent extracellular matrix (atypical spindle cell lipoma morphology); these spindle cell-poor subtypes tend to occur in the hands and feet and may resemble myxoid spindle cell lipoma, except for the presence of nuclear atypia/hyperchromasia and the anatomical location [1, 2, 4, 9, 12]. At the high-cellularity extreme, tumours may be quite cellular, composed of numerous spindle cells showing diffuse, mild to moderate cytonuclear atypia, with easily identified lipoblasts and less extracellular matrix (fibrosarcoma-like lipomatous neoplasm morphology) [1, 2, 13, 12]. A rare finding is heterologous (metaplastic) differentiation, including the presence of smooth muscle, cartilaginous, and/or osseous elements [1, 5].
The tumour cells show variable expression of CD34, S100, and desmin [1, 2]. Weak and/or focal expression of MDM2 or CDK4 can be rarely seen [1, 4, 6]; the combination of MDM2 and CDK4 expression is not encountered [1]. Loss of nuclear RB1 expression is observed in about 50–70% of cases [1, 2, 3, 4].

Immunophenotype. None (immunomarker profile detailed above under Histopathology).

Differential Diagnosis. None.

Cytology. There are few reports, but FNA may show cells similar to those seen histologically [15].

Diagnostic Molecular Pathology. Molecular studies have shown a consistent absence of MDM2 or CDK4 amplification.

Essential and Desirable Diagnostic Criteria. Essential: variable proportions of atypical spindle cells, adipocytes, univacuolated or bivacuolated to multivacuolated lipoblasts, pleomorphic (multinucleated) cells, and a myxoid to collagenous extracellular matrix. Desirable (in selected cases): RB1 expression loss correlating with RB1 deletion; lack of MDM2 or CDK4 amplification.

Grading / Staging. Not clinically relevant.

Prognosis and Prediction. Atypical spindle cell / pleomorphic lipomatous tumour has a low rate of local recurrence (10–15%) for incompletely removed lesions. There is no documented risk for metastasis. Most patients will have an excellent prognosis if the lesion is completely excised [1, 2, 3, 14, 9].

References.
[1] Mariño-Enriquez A, Nascimento AF, Ligon AH, et al. Atypical spindle cell lipomatous tumor: clinicopathologic characterization of 232 cases demonstrating a morphologic spectrum. *Am J Surg Pathol.* 2017;41(2):234–44. PMID:27879515
[2] Creytens D, Mentzel T, Ferdinande L, et al. "Atypical" pleomorphic lipomatous tumor: a clinicopathologic, immunohistochemical and molecular study of 21 cases, emphasizing its relationship to atypical spindle cell lipomatous tumor and suggesting a morphologic spectrum. *Am J Surg Pathol.* 2017;41(11):1443–55. PMID:28877053
[3] Bahadır B, Behzatoğlu K, Hacıhasanoğlu E, et al. Atypical spindle cell/pleomorphic lipomatous tumor: a clinicopathologic, immunohistochemical, and molecular study of 20 cases. *Pathol Int.* 2018;68(10):550–6. PMID:30198097
[4] Creytens D, van Gorp J, Savola S, et al. Atypical spindle cell lipoma: a clinicopathologic, immunohistochemical, and molecular study emphasizing its relationship to classical spindle cell lipoma. *Virchows Arch.* 2014;465(1):97–108. PMID:24659226
[5] Creytens D, Ferdinande L, van Gorp J, et al. Atypical spindle cell lipomatous tumor with benign heterologous (metaplastic) cartilaginous differentiation. *Int J Surg Pathol.* 2019;27(5):521–3. PMID:3070805
[6] Mentzel T, Palmedo G, Kuhnen C. Well-differentiated spindle cell liposarcoma ('atypical spindle cell lipomatous tumor') does not belong to the spectrum of atypical lipomatous tumor but has a close relationship to spindle cell lipoma: clinicopathologic, immunohistochemical, and molecular analysis of six cases. *Mod Pathol.* 2010;23(5):729–36. PMID:20228779
[7] Italiano A, Chamboissiere M, Attias R, et al. Monosomy 7 and absence of 12q amplification in two cases of spindle cell liposarcomas. *Cancer Genet Cytogenet.* 2008;184(2):99–104. PMID:18617058
[8] Agaimy A. Anisometric cell lipoma: insight from a case series and review of the literature on adipocytic neoplasms in survivors of retinoblastoma suggest a role for RB1 loss and possible relationship to fat-predominant ("fat-only") spindle cell lipoma. *Ann Diagn Pathol.* 2017;29:52–6. PMID:28807343
[9] Creytens D, Mentzel T, Ferdinande L, et al. "Fat-rich" (spindle cell-poor) variants of atypical spindle

cell lipomatous tumor show similar morphologic, immunohistochemical and molecular features as "dysplastic lipomas": Are they related lesions? *Am J Surg Pathol.* 2019;43(2):288–9. PMID:30211727
[10] Creytens D, Mentzel T, Ferdinande L, et al. Atypical mitoses are present in otherwise classical pleomorphic lipomas—reply. *Hum Pathol.* 2018;81:300–2. PMID:30084357
[11] Creytens D, Mentzel T, Ferdinande L, et al. Atypical multivacuolated lipoblasts and atypical mitoses are not compatible with the diagnosis of spindle cell/pleomorphic lipoma. *Hum Pathol.* 2018;74:188–9. PMID:29317234
[12] Creytens D. A contemporary review of myxoid adipocytic tumors. *Semin Diagn Pathol.* 2019;36(2):129–41. PMID:30853315
[13] Deyrup AT, Chibon F, Guillou L, et al. Fibrosarcoma-like lipomatous neoplasm: a reappraisal of so-called spindle cell liposarcoma, defining a unique lipomatous tumor unrelated to other liposarcomas. *Am J Surg Pathol.* 2003;27(9):1373–8. PMID:23887155
[14] McCarthy AJ, Chetty R. Tumours composed of fat are no longer a simple diagnosis: an overview of fatty tumours with a spindle cell component. *J Clin Pathol.* 2018;71(6):483–92. PMID:29358476
[15] Yong M, Raza AS, Greaves TS, et al. Fine-needle aspiration of a pleomorphic lipoma of the head and neck: a case report. *Diagn Cytopathol.* 2005;32(2):110–3. PMID:15637670

G.3 Step 1 — Claim Decomposition

Each guideline (GT and prediction) is first decomposed into atomic, independently verifiable claims using the prompt in Appendix A.2.3. Table 15 shows a representative subset of GT claims for the ASPLT case, spanning six knowledge types and eight sections.

GT claim distribution by section (56 total): Histopathology 16, Localization 7, Epidemiology 6, Definition 5, Macroscopic Appearance 5, Essential Diagnostic Criteria 3, Prognosis 3, Related Terminology 3, Clinical Features 2, ICD-O Coding 2, Pathogenesis 2, Cytology 1, Diagnostic Molecular Pathology 1.

GT claim distribution by knowledge type: Factual 42, Diagnostic 5, Prognostic 4, Clinical 2, Mechanistic 2, Differential 1.

Table 16 lists all 56 GT claims with their section, knowledge type, and number of supporting references.

Tables 13 and 14 compare claim distribution between the GT and the GPT-4.1 prediction, by section and knowledge type respectively. Sections absent in the GT but generated by GPT-4.1 are shown in italics.

For the GPT-4.1 prediction, 76 claims are extracted. Notably, the type distribution collapses toward `Factual` (71/76 = 93.4%), with only 4 Prognostic and 1 Differential claim, indicating that the model struggles to replicate the diagnostic and mechanistic reasoning present in the GT.

Section	GT	GPT-4.1	Δ
Histopathology	16	7	−9
Localization	7	3	−4
Epidemiology	6	3	−3
Definition	5	5	0
Macroscopic appearance	5	3	−2
Essential & desirable diagnostic criteria	3	6	+3
Prognosis and prediction	3	4	+1
Related terminology	3	4	+1
Clinical features	2	4	+2
ICD-O coding / ICD-11 coding	2	2	0
Pathogenesis	2	4	+2
Cytology	1	4	+3
Diagnostic molecular pathology	1	3	+2
Differential diagnosis (GT: absent)	0	6	+6
Imaging (GT: absent)	0	4	+4
Immunophenotype (GT: absent)	0	4	+4
Subtype(s) (GT: absent)	0	3	+3
Grading / Staging (GT: absent)	0	3	+3
Spread (GT: absent)	0	3	+3
Etiology (GT: absent)	0	1	+1
Total	**56**	**76**	**+20**

Table 13: GT vs. GPT-4.1 claim counts by section for the ASPLT case. Italicised sections are absent in the GT.

Knowledge Type	GT	GPT-4.1	Δ	Note
Factual	42	71	+29	
Diagnostic	5	0	−5	**complete loss**
Prognostic	4	4	0	
Clinical	2	0	−2	**complete loss**
Mechanistic	2	0	−2	**complete loss**
Differential	1	1	0	
Total	**56**	**76**	**+20**	

Table 14: GT vs. GPT-4.1 claim counts by knowledge type for the ASPLT case. Complete loss of Diagnostic, Clinical, and Mechanistic claim types is highlighted.

ID	Claim Text	Section	Type	Refs
C001	ICD-O code: 3857/0	ICD-O Coding	Factual	0
C002	ICD-11 code: 2E80 & XH4E98	ICD-O Coding	Factual	0
C003	Acceptable synonym: atypical spindle cell lipoma.	Related Terminology	Factual	0
C004	Not recommended: spindle cell liposarcoma.	Related Terminology	Factual	0
C005	Not recommended: fibrosarcoma-like lipomatous neoplasm.	Related Terminology	Factual	0
C006	ASPLT is a benign adipocytic neoplasm.	Definition	Factual	0
C007	ASPLT is characterized by ill-defined tumour margins.	Definition	Factual	0
C008	ASPLT contains variable proportions of mild to moderately atypical spindle cells, adipocytes, lipoblasts, pleomorphic cells, multinucleated giant cells, and a myxoid or collagenous matrix.	Definition	Factual	0
C009	ASPLT has a low tendency for local recurrence if incompletely excised.	Definition	Prognostic	0
C010	Unlike conventional atypical lipomatous tumours, ASPLT has no risk for dedifferentiation.	Definition	Differential	0
C011	The tumour manifests as a persistent or enlarging soft tissue mass, nodule, or swelling.	Clinical Features	Clinical	1
C012	The tumour sometimes manifests with tenderness.	Clinical Features	Clinical	1

Table 15: Representative GT claims for the ASPLT case (C001–C012 shown; full set has 56 claims).

ID	Claim Text	Section	Type	Refs
C001	3857/0 Atypical spindle cell/pleomorphic lipomatous tumour	ICD-O/11	Factual	0
C002	2E80 & XH4E98 Benign lipomatous neoplasm & Spindle cell lipoma	ICD-O/11	Factual	0
C003	Acceptable synonym: atypical spindle cell lipoma	Related Term.	Factual	0
C004	Not recommended: spindle cell liposarcoma	Related Term.	Factual	0
C005	Not recommended: fibrosarcoma-like lipomatous neoplasm	Related Term.	Factual	0
C006	ASPLT is a benign adipocytic neoplasm	Definition	Factual	0
C007	ASPLT is characterized by ill-defined tumour margins	Definition	Factual	0
C008	ASPLT contains variable proportions of mild to moderately atypical spindle cells, adipocytes, lipoblasts, pleomorphic cells, multinucleated giant cells, and a myxoid or collagenous extracellular matrix	Definition	Factual	0
C009	ASPLT has a low tendency for local recurrence if incompletely excised	Definition	Prognostic	0
C010	Unlike conventional atypical lipomatous tumours, ASPLT has no risk for dedifferentiation	Definition	Differential	0
C011	The tumour manifests as a persistent or enlarging soft tissue mass, nodule, or swelling	Clinical	Clinical	1
C012	The tumour sometimes manifests with tenderness	Clinical	Clinical	1
C013	ASPLT arises in the subcutis slightly more frequently than in deep (subfascial) somatic soft tissues	Localization	Factual	0
C014	ASPLT only occasionally arises in intracavitary or visceral locations	Localization	Factual	0
C015	Wide anatomical distribution, predominating in the limbs and limb girdles	Localization	Factual	3
C016	Most common locations: hand and foot and the thigh	Localization	Factual	0
C017	Other common locations: shoulder and buttock, forearm, knee, lower leg, and upper arm	Localization	Factual	0
C018	Less common locations: head and neck, genital area, trunk, and back	Localization	Factual	3
C019	Rare sites: larynx, mediastinum, retroperitoneum, trachea, and appendix	Localization	Factual	1
C020	ASPLT occurs predominantly in middle-aged adults	Epidemiology	Factual	0
C021	Peak incidence in the sixth decade of life	Epidemiology	Factual	0
C022	ASPLT can affect patients of any age	Epidemiology	Factual	3
C023	Cases described in patients aged 6–87 years	Epidemiology	Factual	3
C024	Large majority of patients are >30 years old	Epidemiology	Factual	0
C025	Slight male predominance	Epidemiology	Factual	0
C026	Deletions or losses of 13q14 (including RB1 and flanking genes RCBTB2, DLEU1, IM2B) identified in a substantial subset	Pathogenesis	Mechanistic	6
C027	Monosomy 7 reported in some cases	Pathogenesis	Mechanistic	2
C028	Tumours are unencapsulated	Macroscopic	Factual	0
C029	Tumours show a nodular or multinodular growth pattern	Macroscopic	Factual	0
C030	Tumours demonstrate ill-defined tumour margins	Macroscopic	Factual	0
C031	Tumour size is variable (range: 0.5–28 cm)	Macroscopic	Factual	2
C032	Median tumour size is 5–8.5 cm	Macroscopic	Factual	2
C033	FNA may show cells similar to those seen histologically	Cytology	Diagnostic	1
C034	Molecular studies show consistent absence of MDM2 or CDK4 amplification	Diag. Mol.	Diagnostic	0
C035	Essential: variable proportions of atypical spindle cells, adipocytes, univacuolated or bivacuolated to multivacuolated lipoblasts, pleomorphic cells, and a myxoid to collagenous matrix	Ess. Diag.	Diagnostic	0
C036	In a substantial subset, RB1 expression is lost, correlating with RB1 deletion	Ess. Diag.	Diagnostic	0
C037	Lack of MDM2 or CDK4 amplification is desirable in selected cases	Ess. Diag.	Diagnostic	0
C038	Low rate of local recurrence (10–15%) for incompletely removed lesions	Prognosis	Prognostic	0
C039	No documented risk for metastasis	Prognosis	Prognostic	0
C040	Excellent prognosis if completely excised	Prognosis	Prognostic	5
C041	Wide range of microscopic appearances depending on proportions of spindle cells, adipocytes, lipoblasts, and pleomorphic cells, and extracellular matrix amount	Histopathology	Factual	2
C042	Adipocytic component has predominantly mature morphology, with variation in adipocytic size and shape	Histopathology	Factual	0

Continued on next page

Table 16 continued.

ID	Claim Text	Section	Type	Refs
C043	Patchy, mild to moderate adipocytic atypia with chromatin coarsening, nuclear enlargement, and focal binucleation or multinucleation	Histopathology	Factual	2
C044	Lipoblasts vary from small univacuolated/bivacuolated to larger multivacuolated (pleomorphic)	Histopathology	Factual	0
C045	Bizarre, hyperchromatic, pleomorphic multinucleated cells often scattered within spindle cell or adipocytic components	Histopathology	Factual	0
C046	Mitotic figures often present but mostly scarce	Histopathology	Factual	3
C047	Tumour necrosis is absent	Histopathology	Factual	0
C048	Morphology represents a broad spectrum defined by two morphological extremes	Histopathology	Factual	1
C049	At the paucicellular extreme: few cytologically bland spindle cells with minimal atypia in a prominent extracellular matrix	Histopathology	Factual	0
C050	Spindle cell-poor subtypes tend to occur in the hands and feet and may resemble myxoid spindle cell lipoma, except for nuclear atypia/hyperchromasia and anatomical location	Histopathology	Factual	5
C051	At the cellular extreme: numerous spindle cells with diffuse cytonuclear atypia, easily identified lipoblasts, and less extracellular matrix	Histopathology	Factual	4
C052	Rare finding: heterologous (metaplastic) differentiation (smooth muscle, cartilaginous, or osseous elements)	Histopathology	Factual	2
C053	Tumour cells show variable expression of CD34, S100, and desmin	Histopathology	Factual	2
C054	Weak and/or focal expression of MDM2 or CDK4 can be rarely seen	Histopathology	Factual	3
C055	The combination of MDM2 and CDK4 expression is not encountered	Histopathology	Factual	1
C056	Loss of nuclear RB1 expression is observed in about 50–70% of cases	Histopathology	Factual	4

Table 16: Complete ground-truth claim set for the ASPLT case (all 56 claims).