Supporting Information

Text S1. Full Prompt Used for Large Language Model Evaluation.

"As a specialist in vascular anomalies (including infantile hemangioma and vascular malformations), provide concise answers to all eight clinical questions (CQs) below. Each answer should include: (1) a recommendation, (2) the strength of the recommendation, (3) the level of evidence, and (4) references with links in Vancouver style (list all publications used for items 1–3; limit the sources to articles published between 1980 and 2014 in PubMed, the Cochrane Library, and Ichushi-Web). Refer to the uploaded GRADE guidelines. Classify the strength of recommendation as '1' (strong recommendation) or '2' (weak recommendation [suggestion]); if a determination cannot be made, state 'No recommendation.' Classify the strength of evidence into four levels: A (high), B (moderate), C (low), and D (very low). Provide only one strength of recommendation and one level of evidence for each CQ."

Table S1. Reproducibility across Five Independent Runs.

Model	Language	Correctness complete consistency, n/8 (%)	Recommendation strength complete consistency, n/8 (%)	Evidence level complete consistency, n/8 (%)
ChatGPT 5.1 Thinking	English	5/8 (62.5)	7/8 (87.5)	3/8 (37.5)
	Japanese	6/8 (75.0)	7/8 (87.5)	5/8 (62.5)
Claude Opus 4.1	English	4/8 (50.0)	6/8 (75.0)	3/8 (37.5)
	Japanese	6/8 (75.0)	3/8 (37.5)	4/8 (50.0)
Overall		21/32 (65.6)	23/32 (71.9)	15/32 (46.9)

Note: Complete consistency was defined as identical outputs across all five independent runs within each CQ-model-language unit.

Table S2. Aggregated 32-unit Sensitivity Analysis.

Outcome	ChatGPT 5.1 Thinking		Claude Opus 4.1	
	English	Japanese	English	Japanese
Correctness, median of unit-level medians	5	5	5	5
Majority recommendation-grade concordance, n/8 (%)	5/8 (62.5)	4/8 (50.0)	5/8 (62.5)	5/8 (62.5)
Majority evidence-level concordance, n/8 (%)	1/8 (12.5)	1/8 (12.5)	3/8 (37.5)	2/8 (25.0)
Number of citations, median of unit-level medians	8.5	9.5	5	4
Reference overlap with guideline, median of unit-level medians (%)	22.2	22.6	24.3	12.5
Valid links, median of unit-level medians (%)	100	100	67.5	58.3

Note: For each CQ-model-language unit, the five repeated outputs were aggregated using the median correctness score and majority classification for recommendation-grade and evidence-level concordance.

Table S3. Comparison of Large Language Model-generated Recommendations with the Japanese Clinical Practice Guidelines for Vascular Anomalies 2017 for CQ14 “Is propranolol safe and effective for infantile hemangiomas?”

	Recommendation
Guideline	If administered under careful monitoring, oral propranolol therapy may be the first choice for the treatment of infantile hemangioma.
ChatGPT 5.1 Thinking	For infantile hemangiomas that require systemic treatment (e.g., risk to vision or airway, ulceration, functional impairment, or risk of significant disfigurement), oral propranolol (typically 2-3 mg/kg/day in divided doses) is recommended as first-line therapy, provided that baseline cardiac/respiratory evaluation is performed and appropriate monitoring for hypoglycemia, hypotension, bradycardia, and bronchospasm is ensured.
Claude Opus 4.1	Propranolol is recommended as safe and effective first-line treatment for problematic infantile hemangiomas requiring systemic therapy. Oral propranolol at a dose of 2-3 mg/kg/day demonstrates high efficacy with acceptable safety profile when appropriately monitored.

Note: Both large language model outputs shown here were assigned the maximum correctness rating of 5 on the study’s 5-point scale; this example is illustrative, and aggregate performance metrics and definitions of concordance are reported in the main manuscript.