This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material

S1. LLM-judge prompt (verbatim)

The verbatim grading instruction given to the judge, identical across the subagent and API judging paths:

```
You are a pathology diagnosis grader (LLM-as-judge). Grade how well a model's predicted
diagnosis matches the validated ground-truth diagnosis, calibrated to how expert pathologists grade.

You are given ONE case as JSON: {case_id, model, sex, age, clinical_information, question,
ground_truth, prediction}. Compare 'prediction' to 'ground_truth' (clinical_information and question
are context only).

Reason BEFORE deciding, via three sub-checks:
- entity_match: is the core diagnosis / lineage the same as ground_truth? (e.g. adenocarcinoma,
  leiomyoma, seborrheic keratosis). Tolerate synonyms, abbreviations (BPH = benign prostatic
  hyperplasia), paraphrase, and phrasing like "The most likely diagnosis is X".
- subtype_match: does the subtype/variant match, WHERE ground_truth specifies one? ("n/a" if none).
- grade_qualifier_match: does the grade/stage/key qualifier match, WHERE specified? ("n/a" if none).

RULES (calibrated to the pathologists, who are lenient):
- HEDGED DIFFERENTIALS: if the correct diagnosis is named as the primary/most-likely answer, grade
  on that. A trailing differential list does NOT lower the verdict.
- Grade on clinical substance, not wording.

Assign exactly ONE verdict:
- "Correct": entity_match is yes AND no clinically material contradiction - EVEN IF a subtype/grade/
  qualifier is missing or slightly off, UNLESS that difference would change clinical management (then
  "Partially correct").
- "Partially correct": right general category but a DEFINING feature is wrong/missing such that it
  names a clinically different entity (e.g. wrong tumor grade group that changes management, benign vs
  borderline vs malignant confusion), OR only one of several required components is given.
- "Incorrect": wrong entity, or contradicts ground_truth.
- "Not evaluable": no diagnosis offered / refusal / empty prediction.

Keep the rationale to one short sentence.
```

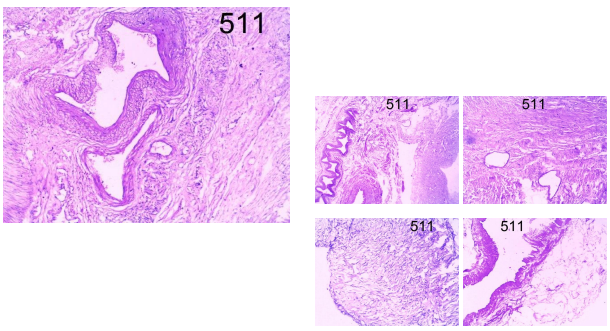
S2. Human grading protocol

The pathologist reviewers used the same four labels and rubric shown in S1 (Correct / Partially correct / Incorrect / Not evaluable). Two reference models (MedGemma and PaiChat) were run on the raw consultation input and every answer was graded by a reviewing pathologist; these n=1372 human grades ground the LLM-judge validation reported in the main paper.

S3. End-to-end example (case PAIX_0689)

Ground truth: Serous cystadenoma. **Accepted diagnoses:** Serous cystadenoma. **Organ:** Ovary; difficulty medium. Standardized no-question, single-image protocol (the referring question is dropped, so any answer leakage is removed).

Model input (system prompt + clinical context)



Representative image (left, large; the one used in the single-image protocol) and 4 further views of the same case.

```
You are an expert pathologist reviewing a histopathology case.
Based on the microscopic image(s) and the clinical context below, give the single most likely histopathological diagnosis.
Be concise and specific; state the diagnosis first, then a brief justification.

Clinical context:
Sex:  Female
Age:  53 years
Clinical Information:  cystic mass measured 11x8x7 cm , no solid component, thin wall with follpian tube .

Diagnosis:
```

Model outputs and LLM-judge review

Model	Output	Verdict	Judge rationale
Claude-Sonnet	Serous cystadenoma (paratubal/ovarian); gland-like spaces lined by flat to cuboidal epithelium within a fibrous strom...	Correct	Primary diagnosis is serous cystadenoma (paratubal/ovarian); matches the accepted diagnosis...
Claude-Haiku	Simple serous cyst	Partially correct	Simple serous cyst is serous but differs from cystadenoma in morphology/classification.
LLaVA-Med	The single most likely histopathological diagnosis for this case is a mature cystic teratoma.	Incorrect	Primary 'mature cystic teratoma' vs accepted 'Serous cystadenoma' — different ovarian cyst e...
Lingshu	Cystadenoma of the ovary. The histopathological features observed in the image, such as the presence of a cystic stru...	Correct	PRIMARY Cystadenoma of the ovary (broader term) encompasses accepted Serous cystadenoma (spe...
MedGemma	Cystic Teratoma The image shows a cystic mass with a thin wall and the presence of follicles, which are characteristi...	Incorrect	Cystic teratoma does not match accepted serous cystadenoma (different cystic entities).
Quilt-LLaVA	The histopathological features observed in this image are consistent with a benign cystic lesion. The absence of a so...	Not evaluable	Prediction explicitly states 'a definitive diagnosis cannot be made' and offers no committed...
PaiChat	**Follicular cyst of the ovary** **Justification:** The image shows a cystic structure with a thin wall and internal...	Incorrect	—

S4. Pairwise model significance

Table 1: Pairwise McNemar tests (binary Correct-vs-not, paired over the cases both runs answered) across all evaluated runs, single- and all-image. *** $p<.001$, ** $p<.01$, * $p<.05$, ns not significant. Full matrix behind the two-tier summary in the main text: the five leaders are mutually ns, each significantly above Quilt-LLaVA and PaiChat.

Model	Claude-Haiku	Claude-Sonnet	LLaVA-Med	Lingshu (all img)	Lingshu	MedGemma (all img)	MedGemma	PaiChat (all img)	PaiChat	Quilt-LLaVA
Claude-Haiku (all img)	***	**	*	ns	ns	ns	ns	ns	*	ns
Claude-Haiku		ns	ns	ns	ns	***	ns	***	***	***
Claude-Sonnet			ns	ns	ns	***	ns	***	***	***
LLaVA-Med				ns	ns	***	ns	***	***	***
Lingshu (all img)					ns	***	ns	***	***	***
Lingshu						**	ns	**	***	***
MedGemma (all img)							**	ns	ns	ns
MedGemma								***	***	***
PaiChat (all img)									ns	ns
PaiChat										ns

S5. All-image vs single-image protocol

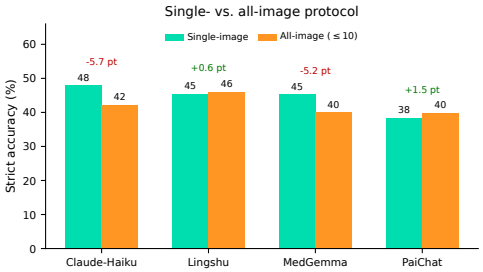


Figure 1: Strict accuracy under the single-image vs all-image ($\leq$10) protocol for the four multi-image models: supplying all case images degrades MedGemma and Claude-Haiku and leaves Lingshu and PaiChat essentially unchanged.

S6. Answer leakage in the referring question

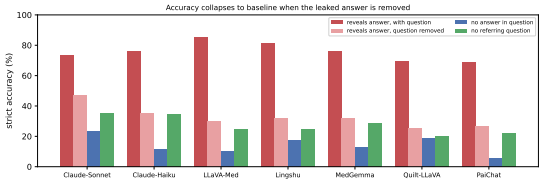


Figure 2: Strict accuracy of the seven single-image baselines by whether the referring question reveals the diagnosis. When the question names a correct/partial diagnosis (261 of 330 question-cases), models score high *with* it but collapse to the no-referring-question baseline once it is removed; the with-question advantage is thus largely referral read-back rather than image reasoning.

S7. Overall and stratified accuracy (difficulty, region)

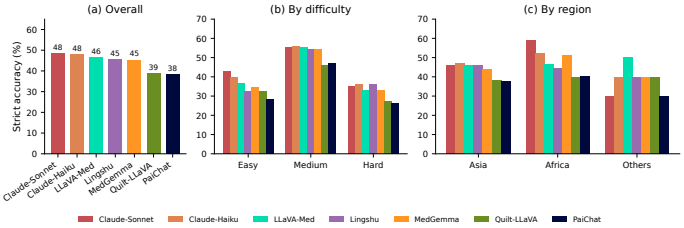


Figure 3: Strict accuracy (% Correct) of the single-image baselines: overall and stratified by case difficulty and world region. Region counts are uneven—Asia (n=528) and Africa (n=148) dominate; Europe and the Americas are merged into "Others" (n=10), which stays noisy.