Supplementary Materials

Neither More Agents nor Different Backbone Models Consistently Reduce Directional Diagnostic Error of Artificial Intelligence Agents in Pulmonary Mucormycosis versus Invasive Pulmonary Aspergillosis

Jiao Xu [1], Jiasong Han [2], Li Li [1] and Xiangdong Mu [1,*]

[1] Department of Respiratory and Critical Care Medicine, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing, China
[2] Department of Electronic Engineering, Tsinghua University, Beijing, China
* Correspondence: mxda02600@btch.edu.cn

This file accompanies the article above and contains Supplementary Methods S1 (complete prompts and analysis rules) and Supplementary Tables S1–S10. Tables S1 and S2 are provided as worksheets in the accompanying file Supplementary_Tables_S1_S2.xlsx because of their length. Throughout, PM denotes pulmonary mucormycosis, IPA invasive pulmonary aspergillosis (strictly invasive spectrum), OTHER another pulmonary diagnosis, and FAIL an unparseable output; all denominators follow the intention-to-evaluate rule, with FAIL retained.

Supplementary Methods S1. Complete Prompts and Analysis Rules

S1.1 Common label vocabulary

```
PM = pulmonary mucormycosis, including disseminated mucormycosis with substantive
pulmonary involvement.
IPA = invasive pulmonary aspergillosis, including disseminated invasive aspergillosis
with substantive pulmonary involvement.
OTHER = any pulmonary diagnosis other than PM or IPA.
Exactly one final label is required. Abstention is not available.
```

S1.2 Common single-agent system prompt

```
You are a physician performing a diagnostic classification task from a case narrative
truncated before the final diagnosis is revealed. Use only information present in the
supplied narrative and any study material explicitly inserted below. Do not assume
that omitted confirmatory results were negative. Select exactly one of PM, IPA, or
OTHER. Explain the evidence briefly, then end with one machine-readable line in the
exact form FINAL_LABEL: PM, FINAL_LABEL: IPA, or FINAL_LABEL: OTHER. Do not add a
second label after the final line.
```

S1.3 Common user prompt template

```
Classify the following pre-confirmation pulmonary case into exactly one category.

CASE_ID: {{case_id}}
TRUNCATED_CASE:
{{case_text}}

{{optional_guideline_primer}}
{{optional_structured_abstraction}}

Return a short rationale followed by exactly one final line:
FINAL_LABEL: PM | IPA | OTHER
```

S1.4 C0–C3 single-agent workflows

C0. Unaided single agent

System prompt: Section S1.2, unchanged. User template: Section S1.3, with guideline primer **absent** and structured abstraction **absent**.

C1. Single agent + primer

System prompt: Section S1.2, unchanged. User template: Section S1.3, with guideline primer **present verbatim as Section S1.10** and structured abstraction **absent**.

C2. Single agent + abstraction

System prompt: Section S1.2, unchanged. User template: Section S1.3, with guideline primer **absent** and structured abstraction **present from Section S1.8**.

C3. Guideline-primed single agent

System prompt: Section S1.2, unchanged. User template: Section S1.3, with guideline primer **present verbatim as Section S1.10** and structured abstraction **present from Section S1.8**.

S1.5 Specialist system prompts used in C4–C7

Infectious-diseases specialist

```
You are the Infectious-diseases specialist in a simulated multidisciplinary
consultation. Read the same truncated case independently. Focus on host factors,
epidemiology, prior antifungal exposure, microbiology, antigen and molecular tests,
and treatment implications. Distinguish evidence that truly separates Mucorales from
Aspergillus from evidence common to invasive mould disease. Do not infer an
unavailable test result. Provide (1) findings supporting PM, (2) findings supporting
IPA, (3) findings supporting another diagnosis, (4) missing discriminating evidence,
and (5) one provisional label: PM, IPA, or OTHER. Do not communicate with the other
specialists and do not claim that agreement among agents is independent evidence.
```

Pulmonary/radiology specialist

```
You are the Pulmonary/radiology specialist in a simulated multidisciplinary
consultation. Read the same truncated case independently. Focus on chronology, lesion
distribution, nodules, consolidation, halo or reversed-halo patterns, cavitation,
pleural findings, vascular invasion, and the likely yield and safety of BAL or tissue
sampling. Treat imaging findings as probabilistic rather than pathognomonic. Provide
(1) findings supporting PM, (2) findings supporting IPA, (3) findings supporting
another diagnosis, (4) missing discriminating evidence, and (5) one provisional label:
PM, IPA, or OTHER. Do not communicate with the other specialists and do not claim that
agreement among agents is independent evidence.
```

Haematology/critical-care specialist

```
You are the Haematology/critical-care specialist in a simulated multidisciplinary
consultation. Read the same truncated case independently. Focus on immune status,
neutropenia, transplantation, corticosteroid or immunosuppressive exposure, clinical
deterioration, feasibility of invasive sampling, and the consequences of delaying
active therapy. Do not convert host risk alone into organism identification. Provide
(1) findings supporting PM, (2) findings supporting IPA, (3) findings supporting
another diagnosis, (4) missing discriminating evidence, and (5) one provisional label:
PM, IPA, or OTHER. Do not communicate with the other specialists and do not claim that
agreement among agents is independent evidence.
```

S1.6 Specialist user prompt template

```
CASE_ID: {{case_id}}
TRUNCATED_CASE:
{{case_text}}

{{optional_guideline_primer}}
{{optional_structured_abstraction}}

Return your evidence assessment and one provisional label.
```

S1.7 Moderator prompt and C4–C7 assembly

```
You are the moderator of a simulated multidisciplinary consultation. You will receive
the truncated case and three specialist assessments generated from the same case.
Their agreement is not independent clinical evidence. Reconcile each claim against the
case text; do not count repeated assertions as additional support. Select exactly one
of PM, IPA, or OTHER. End with exactly one line: FINAL_LABEL: PM, FINAL_LABEL: IPA, or
FINAL_LABEL: OTHER.

CASE_ID: {{case_id}}
TRUNCATED_CASE:
{{case_text}}

{{optional_guideline_primer}}
{{optional_structured_abstraction}}

INFECTIOUS_DISEASES_ASSESSMENT:
{{id_output}}

PULMONARY_RADIOLOGY_ASSESSMENT:
{{pulmonary_output}}
```

```
HAEMATOLOGY_CRITICAL_CARE_ASSESSMENT:
{{haematology_output}}
```

C4, Simulated multi-agent consultation: the three specialist calls and moderator call use the prompts above; primer absent; structured abstraction absent.

C5, Consultation + primer: the three specialist calls and moderator call use the prompts above; primer present; structured abstraction absent.

C6, Consultation + abstraction: the three specialist calls and moderator call use the prompts above; primer absent; structured abstraction present.

C7, Fully specified simulated consultation: the three specialist calls and moderator call use the prompts above; primer present; structured abstraction present.

S1.8 Structured abstraction prompt

```
You are an information extractor, not a diagnostician. Convert the truncated case into
the following fixed fields using only explicit text. Write NOT REPORTED when absent.
Do not infer a final diagnosis and do not add facts.

1. Host factors and immune status
2. Antifungal exposure before the decision point
3. Symptoms and time course
4. Imaging findings and anatomical distribution
5. Microbiology, culture, microscopy, antigen and molecular tests available before
truncation
6. Histopathology available before truncation
7. Treatment response before truncation
8. Missing or pending discriminating investigations

CASE_ID: {{case_id}}
TRUNCATED_CASE:
{{case_text}}
```

The same abstraction was reused for C2, C3, C6, C7, and the rare-disease agent within the same model × case × repetition.

S1.9 Rare-disease agent prompts

Stage 1: phenotype and evidence extraction

```
Extract explicit host factors, pulmonary phenotype, imaging, microbiology/serology,
histopathology, treatment exposure, chronology, and missing discriminating evidence.
Separate positive, negative, and unavailable findings. Do not diagnose. CASE:
{{case_text}}
```

Stage 2: candidate generation

```
Using only the structured evidence below, generate up to five candidate pulmonary
diagnoses. Include uncommon but treatment-critical diseases when compatible. For each
candidate list supporting evidence, opposing evidence, and evidence still required. Do
not use a guideline primer. STRUCTURED_EVIDENCE: {{stage1_output}}
```

Stage 3: self-review and forced label

```
Review the candidates for common-disease anchoring and rare-disease overcalling. Check
every asserted finding against the structured evidence. Then select exactly one
category: PM, IPA, or OTHER. End with exactly one line: FINAL_LABEL: PM, FINAL_LABEL:
IPA, or FINAL_LABEL: OTHER. STRUCTURED_EVIDENCE: {{stage1_output}} CANDIDATES:
{{stage2_output}}
```

S1.10 Verbatim one-page guideline primer

```
Pulmonary mucormycosis (PM) versus invasive pulmonary aspergillosis (IPA): decision
primer

1. Start from overlap. Both diseases occur in severely immunocompromised hosts and may
present with persistent fever, pulmonary nodules, consolidation, infarct-like lesions,
halo sign, cavitation, haemoptysis, or clinical deterioration despite antibacterial
therapy. Neutropenia, haematological malignancy, transplantation, corticosteroid
exposure, and a generic diagnosis of invasive mould infection do not identify the
organism by themselves.

2. Findings that raise concern for PM, without being diagnostic alone, include a
reversed halo sign, pleural effusion, multiple nodules, extensive necrosis or
infarction, rapid progression across tissue planes, diabetes with ketoacidosis, iron
overload, and breakthrough mould infection during voriconazole exposure. Absence of
these findings does not exclude PM.

3. Findings that support IPA include Aspergillus recovered from an appropriate
respiratory or tissue specimen, compatible narrow septate hyphae with predominantly
acute-angle branching, or a positive serum or BAL galactomannan result in a population
in which the assay is validated. Serum galactomannan is most useful in haematological
malignancy and allogeneic HSCT. A single negative result does not exclude IPA,
```

```
particularly during mould-active therapy or in populations with lower assay
sensitivity. Beta-D-glucan is not specific for Aspergillus.

4. Routine galactomannan and beta-D-glucan assays do not directly detect Mucorales.
Negative results may be compatible with PM but are not confirmatory. Positive
galactomannan can reflect Aspergillus infection, cross-reactivity, or mixed infection
and should not be used alone to exclude Mucorales.

5. Seek organism-level evidence promptly. Send adequate tissue or fluid for direct
microscopy or histopathology and fungal culture; where available, add validated
Aspergillus and Mucorales PCR. Broad, ribbon-like, pauciseptate hyphae with irregular
or near-right-angle branching support Mucorales; narrower regularly septate acute-
angle branching supports Aspergillus, but distorted tissue morphology can mislead and
should be combined with culture or molecular identification.

6. Chest CT should be obtained when invasive pulmonary mould disease is suspected.
Bronchoscopy with BAL is generally appropriate for suspected IPA and permits culture,
cytology and BAL galactomannan, but severe hypoxaemia or bleeding risk may preclude
it. BAL yield is lower for peripheral nodules; percutaneous or endobronchial biopsy
should be considered when feasible.

7. Treatment implications are asymmetric. Voriconazole is a primary treatment for IPA
but has no clinically useful activity against Mucorales. Liposomal amphotericin B is a
first-line treatment for mucormycosis; isavuconazole or posaconazole may be
alternatives or step-down/salvage options according to clinical context. Early
surgical evaluation and reversal of modifiable host factors are important in
mucormycosis when feasible.

8. Decision rule for this task. Do not label PM or IPA solely from shared host factors
or nonspecific CT findings. Weigh organism-specific microbiology, histopathology,
molecular evidence, validated antigen results, prior antifungal exposure, and the
direction of radiological progression. When evidence remains incomplete, recognise
that either mould may still be clinically possible, but for this forced task select
the single label best supported by the information visible before confirmation.
```

Source basis for the primer (not part of the prompt presented to models): Patterson TF, et al. IDSA Practice Guidelines for Aspergillosis (2016), https://www.idsociety.org/practice-guideline/aspergillosis/; Cornely OA, et al. Global Guideline for Mucormycosis, *Lancet Infect Dis* (2019), https://pmc.ncbi.nlm.nih.gov/articles/PMC8559573/; Donnelly JP, et al. EORTC/MSGERC Consensus Definitions, *Clin Infect Dis* (2020), https://pmc.ncbi.nlm.nih.gov/articles/PMC7486838/.

S1.11a Unscaffolded frontier-model prompt

```
What is the most likely diagnosis in this case? Choose one: pulmonary mucormycosis,
invasive pulmonary aspergillosis, or another diagnosis.

{{case_text}}
```

No disease-specific system prompt, guideline primer, structured abstraction, specialist role, or moderator was supplied. This 43-case call is separate from C0–C7.

S1.11b Reasoning-mode comparator prompt

```
You are a physician classifying a pulmonary case whose narrative stops before final
diagnostic confirmation. Think through the differential diagnosis carefully using only
the visible case text. Select exactly one of pulmonary mucormycosis, invasive
pulmonary aspergillosis, or another diagnosis. End with exactly one line: FINAL_LABEL:
PM, FINAL_LABEL: IPA, or FINAL_LABEL: OTHER.

CASE:
{{case_text}}
```

This comparator used the DeepSeek V4 Flash backbone with its reasoning mode enabled, a 4,096-token output limit, and no primer, structured abstraction, specialist calls, or moderator. It was run once on the 14 PM cases in the physician subset.

S1.12 Output mapping rules

1. If a valid final machine-readable label is present, use the last exact `FINAL_LABEL` value.

2. Otherwise inspect the final diagnostic commitment, not diagnoses mentioned only in the differential or rationale.

3. Map mucormycosis, Mucorales infection, pulmonary zygomycosis, or a primary diagnosis naming *Rhizopus*, *Mucor*, *Lichtheimia*, *Rhizomucor*, *Cunninghamella*, or another Mucorales genus to PM.

4. Map invasive pulmonary aspergillosis or invasive aspergillosis with substantive pulmonary involvement to IPA. Chronic pulmonary aspergillosis is not IPA and maps to OTHER.

5. Map any other single primary diagnosis to OTHER.

6. If the response refuses the task, supplies no final diagnosis, gives two or more co-equal final categories, is empty after retry, or cannot be resolved under rules 1–5, map to FAIL.

7. A discussion of uncertainty does not cause FAIL when one final category is still clearly selected.

8. Mixed Aspergillus–Mucorales infection has no dedicated class. If the response explicitly names both as a mixed final diagnosis without a dominant category, map to FAIL; if one is explicitly designated the primary diagnosis, map to that category.

9. All automatic FAIL and ambiguous mappings enter manual review. Corrections are applied before analysis, with the original automatic label retained in the audit table.

S1.13 Repetition, failure, and paired-bootstrap specification

Each factorial model × condition was called five times. Temperature was 0 and top-p 1; endpoints did not expose a usable random seed, so these are independent API calls rather than seeded experiments. To avoid duplicate calls, R1 for C3 and C4 was the run of the ten-model comparison (11 August 2026), and R1 for C0 was the run made for the physician subset (16 July 2026); these runs used the same prompts as R2–R5 and were re-parsed with the final mapping rules. R2–R5 for these conditions, and R1–R5 for all other conditions, were new calls made on 18 August 2026. Timeout, HTTP 429, and 5xx errors were retried at most twice with exponential backoff; a still-empty or unparseable response was FAIL. For the two principal contrasts, the PM→IPA indicator was averaged over five calls within case and condition, then the 47 PM cases were resampled with replacement 10,000 times using random seed 20260818. The same sampled case indices were applied to both members of every pair. The resampling stream was generated row-major with Python `random.Random(20260818).randrange(47)` (MT19937); percentile 2.5th and 97.5th quantiles were calculated with NumPy's linear percentile method.

Point estimates and case-level paired bootstrap 95% CIs of the two principal contrasts (percentage points).

Model	C7−C3 point (pp)	C7−C3 95% CI	primer with−without point (pp)	primer with−without 95% CI
DeepSeek V4 Flash	+23.0	+17.4 to +28.5	-22.2	-24.4 to -20.1
GLM-5.2	+9.4	+5.5 to +13.6	-5.4	-6.8 to -4.0
Claude Opus 5	+5.5	+2.6 to +9.4	-2.3	-3.7 to -0.9

S1.14 Source-recognition probe

Each of the three factorial backbones (DeepSeek V4 Flash, GLM-5.2, and Claude Opus 5) received each of the 140 truncated narratives once, in a call separate from all diagnostic runs, with the prompt below followed by the truncated narrative that the agents had received. No candidate labels, guideline primer, structured abstraction, or reference diagnosis was supplied (420 calls in total).

```
You are conducting a bibliographic recognition check. Read the supplied truncated
clinical case text. Without searching external sources, state whether you recognize
the source article. Return JSON only with these fields: recognized, title,
first_author, journal, year, pmid. If the article is not recognized, set all
bibliographic fields to null. Do not infer or guess a source solely from the
diagnosis.
```

Scoring. A response was classified as NO_CLAIM when recognition was not claimed. When recognition was claimed, the stated PubMed identifier (PMID) was resolved in

PubMed and the retrieved record was checked against the stated title, first author, journal, and year. The response was classified as CORRECT when the stated publication was the source article of the narrative (identical PMID), and as WRONG_CLAIM otherwise. Verdicts were fixed before the diagnostic records were re-analysed.

Subsets and analysis. The model-specific subset excluded the mucormycosis cases classified as CORRECT for that model. The strict common subset excluded the cases classified as CORRECT for any of the three models, and the most conservative common subset retained only the cases classified as NO_CLAIM by all three. No diagnostic run was repeated: PM→IPA counts were recomputed from the existing records with the definitions of Section S1.13 and the same case-paired bootstrap (10,000 resamples, seed 20260818), using the subset size as the denominator.

Supplementary Tables

Table S1. Complete confusion counts by model, agent, and repetition. For each of 151 model × agent × repetition rows (three factorial backbones × eight agents × five runs; seven further backbones × two agents × one run; rare-disease agent, three backbones × five runs; unscaffolded frontier model, 43 cases; reasoning-mode comparator, 14 cases), the table gives the number of cases of each reference class (PM, IPA, OTHER) assigned to each output label (PM, IPA, OTHER, FAIL), that is, three reference classes by four output categories: a 3 × 3 confusion matrix with a separate column for FAIL. Provided as sheet Table_S1 of Supplementary_Tables_S1_S2.xlsx.

Table S2. Per-class and aggregate metrics for the same 151 rows, computed from Table S1: precision, recall, and F1 for PM, IPA, and OTHER; accuracy; balanced accuracy; macro-F1; and the number of FAIL outputs. FAIL outputs count against recall and are not counted as a prediction of any class. The reasoning-mode comparator was evaluated on the 14 mucormycosis cases of the physician subset only; for this row the three-class metrics (precision, IPA and OTHER metrics, accuracy, balanced accuracy, and macro-F1) are not applicable and are marked NA, and only mucormycosis recall and FAIL are reported. Provided as sheet Table_S2 of Supplementary_Tables_S1_S2.xlsx.

Table S3. Evidence status at truncation for the 47 mucormycosis cases. Tiers were assigned from the truncated text by two reviewers blinded to all model output; they agreed before adjudication in 39/47 cases (83.0%), and adjudication did not alter the tier totals. For the stratified analyses (Tables S7 and S8), the last four tiers were combined into "any discriminating evidence present".

Evidence status at truncation	*n*	% of 47
No discriminating evidence	21	44.7
Mucorales-leaning evidence	10	21.3
Aspergillus-leaning evidence	9	19.1
Both types present	4	8.5
Other discriminating evidence	3	6.4
Any discriminating evidence present (last four tiers)	26	55.3

Table S4A. Physician reference subset: agreement and complete confusion matrices on the stratified random subset of 43 cases (14 PM, 14 IPA, 15 OTHER). The first four rows of the agreement table involve independent diagnosis from pre-confirmation truncated text; the last row is the agreement between the two reviewers in the dataset review of Section 2.1 of the article (143 cases), in which the final confirmatory information of the source article was visible, listed for contrast. CI, case-level bootstrap 95% confidence interval.

Panel A. Agreement

Comparison	Agreement	Cohen's κ	95% CI
Physician 1 vs reference diagnosis	24/43 (55.8%)	0.339	0.121–0.545
Physician 2 vs reference diagnosis	25/43 (58.1%)	0.373	0.149–0.579
Unscaffolded Claude Opus 5 vs reference diagnosis	25/43 (58.1%)	0.371	0.139–0.581
Physician 1 vs physician 2	39/43 (90.7%)	0.856	0.708–0.965
Dataset review (confirmatory information visible)	140/143 (97.9%)	0.969	0.926–1.000

Panel B. Confusion matrices (rows: reference diagnosis, or physician 1 in the last block; columns: the second rater's label). No FAIL output occurred in this subset.

Comparison	Row label	PM	IPA	OTHER	Total
Reference diagnosis × physician 1	PM	8	1	5	14
	IPA	2	11	1	14
	OTHER	2	8	5	15
Reference diagnosis × physician 2	PM	8	1	5	14
	IPA	1	11	2	14
	OTHER	2	7	6	15
Reference diagnosis × unscaffolded Claude Opus 5	PM	8	1	5	14
	IPA	0	10	4	14
	OTHER	1	7	7	15
Physician 1 × physician 2	PM	11	0	1	12
	IPA	0	18	2	20
	OTHER	0	1	10	11

Table S4B. Manual audit of 378 output mappings (summary). All automatic FAIL and ambiguous mappings were reviewed; the remaining outputs were sampled with stratification by model, agent, and automatic label. Mapping rules are given in Section S1.12.

Audit component	Result
Outputs reviewed	378 (360 stratified sample; 18 FAIL or ambiguous)
Agreement between the two reviewers	376/378 (99.5%)
Automatic mapping identical to the final manual label	375/378 (99.2%)
Mappings corrected before analysis	3 (0.8%)
Corrections involving a PM output mapped to IPA	0

Table S5. Model versions, endpoints, access dates, and decoding settings. All calls used temperature 0 and top-p 1; no endpoint exposed a usable random seed, so repeated runs are independent calls rather than seeded replicates. Timeouts, HTTP 429, and 5xx errors were retried at most twice with exponential backoff. Panel A gives the model versions and settings, and Panel B the runs of each experiment. Access dates are given by experiment: the physician-subset runs, including R1 of the unaided single agent (Methods S1.13), on 16 July 2026; the ten-model comparison on 11 August 2026; the factorial experiment on 18 August 2026; and the source-recognition probe (Section S1.14) on 24 September 2026; for the three factorial backbones, the four dates in the table are those of these four experiments, in this order. For DeepSeek V4 Flash, the provider's release changed between 16 July and 11 August 2026: the physician-subset runs, including R1 of the unaided single agent, used the earlier release, and all later calls the later release. R1 of the unaided single agent fell within the range of R2–R5 on every endpoint (PM→IPA 14 vs 13–16; PM→PM 15 vs 15–16; PM→OTHER 18 vs 15–19; correct labels 74 vs 74–75 of 140).

Panel A. Model versions, endpoints, access dates, and maximum output tokens

Model	Version identifier	Provider and endpoint	Access dates	Maximum output tokens
DeepSeek V4 Flash	deepseek-v4-flash-2026-06-15 (16 July); deepseek-v4-flash-0731 (from 11 August)	DeepSeek official API; https://api.deepseek.com/v1	2026-07-16; 2026-08-11; 2026-08-18; 2026-09-24	2048 final; 3072 specialist/abstraction; 4096 reasoning baseline
Kimi K3	kimi-k3-2026-06-30	Moonshot AI official API; https://api.moonshot.cn/v1	2026-08-11	2048 final; 3072 specialist
Gemini 3.5 Flash	gemini-3.5-flash-002	Google Vertex AI; us-central1-aiplatform.googleapis.com	2026-08-11	2048 final; 3072 specialist
MiniMax M3	minimax-m3-2026-06-20	MiniMax official API; https://api.minimax.chat/v1	2026-08-11	2048 final; 3072 specialist
GLM-5.2	glm-5.2-2026-06-28	Zhipu AI official API; https://open.bigmodel.cn/api/paas/v4	2026-07-16; 2026-08-11; 2026-08-18; 2026-09-24	2048 final; 3072 specialist/abstraction
GLM-5	glm-5-2026-03-10	Zhipu AI official API; https://open.bigmodel.cn/api/paas/v4	2026-08-11	2048 final; 3072 specialist
GPT-5.5	gpt-5.5-2026-06-15	OpenAI Responses API; https://api.openai.com/v1/responses	2026-08-11	2048 final; 3072 specialist
Seed2.1 Pro	doubao-seed-2-1-pro-260615	Volcengine Ark; https://ark.cn-beijing.volces.com/api/v3	2026-08-11	2048 final; 3072 specialist
Qwen3.8-Max	qwen3.8-max-2026-06-18	Alibaba Cloud DashScope compatible API; https://dashscope.aliyuncs.com/compatible-mode/v1	2026-08-11	2048 final; 3072 specialist
Claude Opus 5	claude-opus-5-20260625	Anthropic official API; https://api.anthropic.com/v1/messages	2026-07-16; 2026-08-11; 2026-08-18; 2026-09-24	2048 final; 3072 specialist/abstraction

Panel B. Runs of each experiment. Version identifiers as in Panel A; for DeepSeek V4 Flash, 06-15 and 0731 denote deepseek-v4-flash-2026-06-15 and deepseek-v4-flash-0731. Physician-subset cases:

C0, all 140 cases; reasoning-mode comparator, the 14 mucormycosis cases of the physician subset; unscaffolded call, the 43-case subset. R, repetition.

Experiment	Date	Backbone models	Configurations	Cases	Runs	Reused runs
Physician subset	2026-07-16	DeepSeek V4 Flash (06-15), GLM-5.2, Claude Opus 5	C0; reasoning-mode comparator (DeepSeek V4 Flash); unscaffolded call (Claude Opus 5)	140; 14; 43	1	—
Ten-model comparison	2026-08-11	All ten (DeepSeek V4 Flash, 0731)	C3, C4	140	1	—
Factorial experiment	2026-08-18	DeepSeek V4 Flash (0731), GLM-5.2, Claude Opus 5	C0–C7; rare-disease agent	140	5	R1 of C3 and C4 from the ten-model comparison; R1 of C0 from the physician subset
Source-recognition probe	2026-09-24	DeepSeek V4 Flash (0731), GLM-5.2, Claude Opus 5	Recognition prompt (Methods S1.14)	140	1	—

Table S6. Completed TRIPOD-LLM reporting checklist. Study type under TRIPOD-LLM: large language model (LLM) evaluation (E) in a healthcare setting (H); LLM task: classification (C). Items restricted to LLM methods (M) or de novo LLM development (D), and items restricted to generative task types (QA, IR, DG, SS, MT), are marked *Not applicable* with the reason given. Section, table, and figure numbers refer to the article.

Section	Item	Checklist item (abridged)	Reported in
Title	1	Identify the study as developing, fine-tuning, and/or evaluating an LLM, specifying task, target population, and outcome	Title
Abstract	2	See TRIPOD-LLM for Abstracts	Abstract
Introduction	3a	Healthcare context / use case and rationale, with references to existing approaches and models	Section 1; Table 1
	3b	Target population and intended use in the care pathway, including intended users	Section 1; Section 4.5
	4	Study objectives, including whether development, fine-tuning, or validation	Section 1, final paragraph
Methods — Data	5a	Sources of data for training, tuning, and/or evaluation, and rationale	Section 2.1. No training or tuning data: pre-trained commercial LLMs were evaluated as supplied
	5b	Relevant data points and quantitative/qualitative description of their distribution and descriptors	Section 2.2; Table 2; Section 4.6 (publication years 2005–2024); Section 4.8 (English-language published case reports). Country of origin of the source reports was not recorded (Note 1)
	5c	Date of oldest and newest text item in development and evaluation datasets	Section 4.6: source case reports published between 2005 and 2024. No training or fine-tuning data were used
	5d	Data pre-processing and quality checking	Sections 2.1, 2.2, 2.7; Figure 1
	5e	Handling of missing and imbalanced data; reasons for omitting data	Sections 2.1 (exclusions with counts), 2.3 (FAIL retained in denominator)
Analytical methods	6a	LLM name, version, and last date of training	Sections 2.4, 2.5; Table S5. Training end dates are not published for all commercial endpoints; stratification by knowledge cut-off is described in Section 2.7
	6b	LLM development process, architecture, training, fine-tuning, alignment	*Not applicable*: restricted to M/D; this study evaluates existing commercial LLMs and performed no development or fine-tuning
	6c	How text was generated, prompt engineering, inference settings	Section 2.4 (temperature 0, top-p 1, no usable seed exposed); Methods S1 (all prompts verbatim); Methods S1.13; Table S5
	6d	Initial and post-processed LLM output	Section 2.3 (label space); Sections 2.7 and 3.7 (output mapping and its audit); Methods S1.12; Table S4B
	6e	Classification details; how probabilities were determined and thresholds identified	Section 2.3. Forced three-class labels were elicited directly; no probabilities or thresholds were used
LLM output	7a	Metrics capturing the quality of generative outputs	*Not applicable*: restricted to generative tasks (QA, IR, DG, SS, MT); this study's task is classification
	7b	Relevance of outcome metrics to the downstream deployment task; correlation with human evaluation	Sections 2.3, 4.4; Table 8
	7c	Definition of the outcome, how predictions were calculated, date of inference for closed-source LLMs, evaluation metrics	Sections 2.3, 2.4; Table S5
	7d	Qualifications of assessors, instructions given, and inter-assessor agreement where interpretation is subjective	Sections 2.1, 2.2, 2.6, 3.6; Table 8; Table S4A
	7e	How performance was compared with other LLMs, humans, and benchmarks	Sections 2.5, 2.6, 3.2, 3.6; Tables 4, 7, 8

Section	Item	Checklist item (abridged)	Reported in
Annotation	8a	How text was labelled, including annotation guidelines with examples	Sections 2.1, 2.2; Methods S1
	8b	Number of annotators, proportion of data double-annotated, and inter-annotator agreement	Sections 2.1 (143/143 double-annotated; $\kappa = 0.969$), 2.2 (47/47; 39/47 before adjudication), 2.6 (43/43; $\kappa = 0.856$); Table 8; Tables S3, S4A
	8c	Background and experience of the annotators	Sections 2.1, 2.6
Prompting	9a	Prompt design, curation, and selection	Section 2.4; Methods S1
	9b	Data used to develop the prompts	Section 2.4 (primer compiled from EORTC/MSGERC, ECMM, and IDSA sources [20,23,26]); Methods S1.10
Summarization	10	Preprocessing before summarization	*Not applicable*: restricted to summarization tasks
Instruction tuning / alignment	11	Instructions, data, and interface used for evaluation	*Not applicable*: restricted to M/D; no instruction tuning or alignment was performed
Compute	12	Compute or proxies required to carry out the methods	Note 2
Ethical approval	13	Institutional review board or ethics committee; consent or waiver	Institutional Review Board Statement; Informed Consent Statement
Open science	14a	Source of funding and role of funders	Funding
	14b	Conflicts of interest and financial disclosures	Conflicts of Interest
	14c	Where the study protocol can be accessed, or state that none was prepared	Section 2.8: no protocol was prepared in advance
	14d	Registration information, or state that the study was not registered	Section 2.8: the study was not registered
	14e	Availability of the study data	Data Availability Statement; Tables S1–S10
	14f	Availability of code to reproduce the study results	Data Availability Statement
Public involvement	15	Patient and public involvement, or state no involvement	Section 2.8: no patient or public involvement
Results — Participants	16a	Flow of text/data through the study, with numbers by outcome	Figure 1; Sections 2.1, 3.1
	16b	Characteristics overall and by source/setting and development/evaluation split, including key dates and sample size	Table 2; Sections 2.1, 3.1, 4.6 (publication years 2005–2024). Evaluation-only study; no development/evaluation split
	16c	Distribution of important clinical variables between development and evaluation data	*Not applicable*: evaluation-only study with no development dataset
	16d	Number of participants and outcome events in each analysis	Sections 2.3 (intention-to-evaluate denominators), 3.1–3.7; Tables 4–8
Performance	17	LLM performance according to pre-specified metrics and/or human evaluation	Sections 3.2–3.7; Tables 4–8; Figures 2 and 3; Tables S1, S2
LLM updating	18	Results of any LLM updating	*Not applicable*: no model updating was performed
Discussion	19a	Overall interpretation of the main results, including fairness	Sections 4.1–4.4
	19b	Limitations and their effects on bias, statistical uncertainty, and generalizability	Section 4.8
	19c	Known challenges in using the data for the task and domain (representation, missingness, harmonisation, bias)	Sections 3.1, 4.5, 4.6; Section 4.7 and Table S9 (training-data exposure); Section 4.8
	19d	Intended use of the implementation evaluated: input, end-user, level of autonomy/human oversight	Section 4.5 (Table 9); Section 4.8
	19e	How poor-quality or unavailable input data should be assessed and handled	Sections 3.1, 4.5 (Table 9, fields 2 and 4); Box 1
	19f	Whether users must interact in handling input data, and the expertise required	Section 4.5
	19g	Next steps for future research, with a view to applicability and generalizability	Sections 4.5, 4.8, 5

Note 1 (item 5b, country of origin of the source case reports). The publication years of the source articles (2005–2024) are reported in Section 4.6 of the article. Country of origin was not retained in the RareArena metadata used for this study and is therefore not reported.

Note 2 (item 12, compute). All inference was performed through the application programming interfaces of hosted commercial endpoints; no local training, fine-tuning, or model hosting was performed. The study comprised 20,917 final diagnostic outputs (Table S1), each requiring one to five API calls depending on the agent (Sections S1.4–

S1.9), and 420 source-recognition probe calls (Section S1.14). Total token usage and wall-clock time are not reported.

Table S7. Content-matched consultation versus single agent, stratified by evidence status. Mean PM→IPA cases across five runs [range]; percentages use the stratum size as denominator. "None", no discriminating evidence at truncation ($n = 21$); "Any", any discriminating evidence present ($n = 26$). This stratification is exploratory and descriptive; no interaction was tested.

Model	Evidence status	*n*	Guideline-primed single agent	Fully specified consultation	Difference
DeepSeek V4 Flash	None	21	2.4 [2–3] (11.4%)	9.6 [9–11] (45.7%)	+34.3 pp
	Any	26	2.0 [2–2] (7.7%)	5.6 [5–6] (21.5%)	+13.8 pp
GLM-5.2	None	21	3.0 [2–4] (14.3%)	6.4 [5–8] (30.5%)	+16.2 pp
	Any	26	3.0 [3–3] (11.5%)	4.0 [4–4] (15.4%)	+3.8 pp
Claude Opus 5	None	21	2.0 [1–3] (9.5%)	4.2 [3–5] (20.0%)	+10.5 pp
	Any	26	2.0 [2–2] (7.7%)	2.4 [2–3] (9.2%)	+1.5 pp

Table S8. Marginal effect of the guideline primer, stratified by evidence status. PM→IPA as a percentage of the stratum, averaged over the matched agent pairs within each backbone. This stratification is exploratory and descriptive; no interaction was tested.

Model	Evidence status	*n*	Without primer	With primer	Difference
DeepSeek V4 Flash	None	21	50.0%	31.9%	−18.1 pp
	Any	26	40.0%	14.4%	−25.6 pp
GLM-5.2	None	21	26.2%	22.1%	−4.0 pp
	Any	26	23.1%	16.5%	−6.5 pp
Claude Opus 5	None	21	13.8%	11.9%	−1.9 pp
	Any	26	14.4%	11.7%	−2.7 pp

Table S9. Source-recognition probe and contamination-resistant sensitivity analysis. Recognition was assessed once per case and model (Section S1.14). CORRECT, the stated publication was the source article; WRONG_CLAIM, recognition was claimed but the stated publication was another article; NO_CLAIM, recognition was not claimed. Of the 30 wrong attributions, 25 named a PubMed article outside the case set and 5 named the source article of another case in the set. Fifteen mucormycosis cases were recognised by at least one model. PM→IPA, pulmonary mucormycosis labelled as invasive pulmonary aspergillosis; pp, percentage points.

Panel A. Recognition of the source publication

Model	Cases	*n*	CORRECT	WRONG_CLAIM	NO_CLAIM
DeepSeek V4 Flash	All	140	12 (8.6%)	12 (8.6%)	116 (82.9%)
	PM	47	4 (8.5%)	3 (6.4%)	40 (85.1%)
GLM-5.2	All	140	16 (11.4%)	10 (7.1%)	114 (81.4%)
	PM	47	5 (10.6%)	3 (6.4%)	39 (83.0%)
Claude Opus 5	All	140	21 (15.0%)	8 (5.7%)	111 (79.3%)
	PM	47	7 (14.9%)	2 (4.3%)	38 (80.9%)

Panel B. PM→IPA by agent in the full set and in the model-specific unrecognised subset (mean number of cases over five runs [range])

Agent	DeepSeek V4 Flash, all (47)	DeepSeek V4 Flash, subset (43)	GLM-5.2, all (47)	GLM-5.2, subset (42)	Claude Opus 5, all (47)	Claude Opus 5, subset (40)
C0	14.4 [13–16]	12.8 [11–15]	10.2 [9–11]	9.0 [8–10]	6.0 [5–7]	5.2 [4–6]
C1	7.8 [7–9]	6.6 [4–8]	7.0 [6–8]	6.2 [4–8]	4.0 [3–5]	3.4 [3–4]
C2	10.2 [9–11]	9.0 [8–10]	8.4 [8–9]	7.6 [5–9]	5.0 [4–6]	4.4 [4–6]
C3	4.4 [4–5]	4.0 [3–5]	6.0 [5–7]	5.4 [4–7]	4.0 [3–5]	3.4 [2–5]
C4	34.6 [33–36]	31.2 [30–33]	14.0 [12–16]	12.6 [9–15]	8.0 [5–10]	6.8 [3–10]
C5	14.4 [13–16]	12.8 [11–14]	12.4 [11–14]	11.0 [10–13]	7.6 [7–9]	6.8 [6–8]
C6	24.4 [23–26]	22.0 [20–24]	13.4 [12–15]	12.6 [11–15]	7.6 [7–9]	6.8 [6–8]
C7	15.2 [14–17]	13.6 [12–15]	10.4 [9–12]	9.8 [9–11]	6.6 [6–8]	5.8 [5–6]

Panel C. Principal contrasts in the full set and in the model-specific unrecognised subset (pp, case-paired bootstrap 95% CI)

Model	Guideline-primer effect, full set	Guideline-primer effect, subset	Content-matched consultation cost (C7 − C3), full set	Content-matched consultation cost (C7 − C3), subset
DeepSeek V4 Flash	−22.2 (−24.4 to −20.1)	−22.1 (−24.2 to −19.9)	+23.0 (+17.4 to +28.5)	+22.3 (+16.7 to +27.9)
GLM-5.2	−5.4 (−6.8 to −4.0)	−5.6 (−7.1 to −4.0)	+9.4 (+5.5 to +13.6)	+10.5 (+6.2 to +15.2)
Claude Opus 5	−2.3 (−3.7 to −0.9)	−2.4 (−3.9 to −0.8)	+5.5 (+2.6 to +9.4)	+6.0 (+2.5 to +10.5)

Panel D. Principal contrasts in the two common subsets (pp, case-paired bootstrap 95% CI)

Model	Subset	*n*	Guideline-primer effect	Content-matched consultation cost (C7 − C3)
DeepSeek V4 Flash	Strict: recognised by no model	32	−21.1 (−23.4 to −18.8)	+23.1 (+16.9 to +29.4)
GLM-5.2	Strict: recognised by no model	32	−5.5 (−7.3 to −3.6)	+12.5 (+7.5 to +18.1)
Claude Opus 5	Strict: recognised by no model	32	−2.2 (−3.8 to −0.6)	+6.2 (+1.9 to +11.2)
DeepSeek V4 Flash	Most conservative: no claim by any model	25	−21.0 (−23.8 to −18.4)	+23.2 (+16.8 to +30.4)
GLM-5.2	Most conservative: no claim by any model	25	−5.4 (−7.6 to −3.2)	+13.6 (+7.2 to +20.0)
Claude Opus 5	Most conservative: no claim by any model	25	−3.2 (−5.0 to −1.2)	+3.2 (0.0 to +7.2)

Table S10. Mean factor contrasts for all mucormycosis endpoints and for over-diagnosis of mucormycosis in the factorial experiment, computed from Table S1. Each value is the difference in the mean number of cases across five runs, averaged over the four matched condition pairs within each backbone (as in Table 5, Panel B): consultation minus single agent, and with minus without the primer or structured abstraction; the pair-level differences can be computed from Table S1. Across the 12 matched comparisons (four per backbone), PM→IPA was higher with the consultation structure in all 12 and lower with the guideline primer in all 12. PM→IPA, PM→PM, and PM→OTHER are counts among the 47 mucormycosis cases; non-PM→PM is the count among the 93 non-mucormycosis cases. Because the three backbones were not selected randomly or prospectively, these contrasts are conditional on them. PM, pulmonary mucormycosis; IPA, invasive pulmonary aspergillosis.

Element	Backbone	PM→IPA	PM→PM (recall)	PM→OTHER	Non-PM→PM
Consultation structure	DeepSeek V4 Flash	+12.95	−5.10	−7.85	+0.55
	GLM-5.2	+4.65	−1.90	−2.75	+0.80
	Claude Opus 5	+2.70	−0.75	−1.90	+1.40
Guideline primer	DeepSeek V4 Flash	−10.45	0.00	+10.45	−0.95
	GLM-5.2	−2.55	0.00	+2.55	−0.70
	Claude Opus 5	−1.10	−0.15	+1.20	−0.10
Structured abstraction	DeepSeek V4 Flash	−4.25	−0.30	+4.55	+0.25
	GLM-5.2	−1.35	+0.40	+0.95	+0.50
	Claude Opus 5	−0.60	+0.05	+0.50	+0.60