Supplementary Information

MDT-guided language-model reasoning for spinal infection diagnosis

Supplementary study details

Frozen implementation and formal evaluation

The MDT-guided workflow architecture, prompts, output-parsing rule and model deployment identities were fixed before formal evaluation. All model-by-centre-by-decision comparisons used the same prespecified configuration.

For infection identification, the workflow combines a focused review (minimal-label assessment) with a full-evidence review (complete-record assessment) using a deterministic union rule. For tuberculous-aetiology differentiation, it uses separate clinical-course, laboratory and imaging reviews, followed by conflict review and coordination. The final class follows the frozen output rule. The response score is not a calibrated probability and was excluded from probability-based performance and calibration analyses.

Paired analysis

Each contrast compares the MDT-guided workflow with the direct-label, minimal zero-shot baseline for the same patient record, model, centre and clinical decision. Confidence intervals use 10,000 outcome-stratified paired patient-bootstrap replicates. Exact two-sided McNemar tests are reported as descriptive cellwise checks without a multiplicity-adjusted testing hierarchy. The Centre B infection-identification cohort was disease enriched and included ten non-infection records.

Reader study

Five clinicians completed the same two-session, case-arm crossover schedule. In each session, every clinician assessed 120 cases with fixed MDT-guided AI advice and 120 without advice; exposure was reversed after a six-week washout. The primary reader-study estimand was the available-observation-weighted paired accuracy difference for the observed reader panel. Table S5 reports both a fixed-reader case bootstrap and a reader–case bootstrap; Table S6 reports paired decision transitions.

Infection-identification component decomposition

The infection-identification workflow combines a focused review (minimal-label assessment) and a full-evidence review (complete-record assessment) using a deterministic union rule. The component analysis in Fig. 4 was reconstructed from the completed frozen workflow traces. It compares the minimal-label component, complete-record component and union-rule output in the six model–centre settings. This decomposition is limited to infection identification and does not estimate individual components of the tuberculous-aetiology workflow.

Table S1: Model deployments and fixed inference configuration.

Model	Documentation	Shared evaluation configuration
Gemma 4 31B IT	Technical report and model card	Local deployment; temperature 0; fixed structured schema
Qwen3.6 27B	Model card	Local deployment; temperature 0; fixed structured schema
DeepSeek V4 Flash	Technical report and model card	Local deployment; temperature 0; fixed structured schema

Full citations to the technical reports and model cards appear in the main manuscript. The same model served every role within a model-case workflow.

Table S2: Complete patient-level performance by model, centre and clinical decision.

Centre	Decision	Model	n	Accuracy, %	Balanced accuracy, %	Sensitivity, %	Specificity, %	Positive-class F1, %
Centre A	Infection	Gemma 4 31B IT	325	91.7/92.3	87.6/88.6	77.3/79.4	97.8/97.8	84.7/86.0
Centre A	Infection	Qwen3.6 27B	325	88.3/90.8	81.3/85.7	63.9/73.2	98.7/98.2	76.5/82.6
Centre A	Infection	DeepSeek V4 Flash	325	86.5/89.5	80.3/85.7	64.9/76.3	95.6/95.2	74.1/81.3
Centre A	Aetiology	Gemma 4 31B IT	97	84.5/80.4	81.4/78.9	76.5/76.5	86.3/81.3	63.4/57.8
Centre A	Aetiology	Qwen3.6 27B	97	83.5/74.2	83.1/77.4	82.4/82.4	83.8/72.5	63.6/52.8
Centre A	Aetiology	DeepSeek V4 Flash	97	79.4/61.9	78.2/69.9	76.5/82.4	80.0/57.5	56.5/43.1
Centre B	Infection	Gemma 4 31B IT	89	67.4/68.5	81.6/82.3	63.3/64.6	100/100	77.5/78.5
Centre B	Infection	Qwen3.6 27B	89	60.7/66.3	77.8/81.0	55.7/62.0	100/100	71.5/76.6
Centre B	Infection	DeepSeek V4 Flash	89	62.9/64.0	79.1/79.7	58.2/59.5	100/100	73.6/74.6
Centre B	Aetiology	Gemma 4 31B IT	79	68.4/54.4	67.7/56.6	66.7/60.0	68.8/53.1	44.4/33.3
Centre B	Aetiology	Qwen3.6 27B	79	62.0/51.9	66.4/65.2	73.3/86.7	59.4/43.8	42.3/40.6
Centre B	Aetiology	DeepSeek V4 Flash	79	62.0/55.7	68.9/65.0	80.0/80.0	57.8/50.0	44.4/40.7

Values are reported as direct-label baseline / MDT-guided workflow and are rounded to three significant digits. Patient-level accuracy and balanced accuracy are primary endpoints. Aetiology denotes tuberculous versus non-tuberculous aetiology among reference-standard spinal infections.

Table S3: Discordant paired decisions for tuberculous-aetiology differentiation.

Centre	Model	n	Baseline incorrect → MDT-guided correct	Baseline correct → MDT-guided incorrect
Centre A	Gemma 4 31B IT	97	4	8
Centre A	Qwen3.6 27B	97	3	12
Centre A	DeepSeek V4 Flash	97	7	24
Centre B	Gemma 4 31B IT	79	9	20
Centre B	Qwen3.6 27B	79	8	16
Centre B	DeepSeek V4 Flash	79	9	14

Discordant transitions are paired within the same patient record and model. They characterize the direction of model decision changes for tuberculous-aetiology differentiation.

Table S4: Paired contrasts between the MDT-guided workflow and direct-label baseline for the primary endpoints.

Centre	Decision	Model	n	Δ accuracy, pp (95% CI)	Δ balanced accuracy, pp (95% CI)	McNemar P
Centre A	Infection	Gemma 4 31B IT	325	0.62 (0.00 to 1.54)	1.03 (0.00 to 2.58)	0.500
Centre A	Infection	Qwen3.6 27B	325	2.46 (0.92 to 4.31)	4.42 (1.84 to 7.51)	0.021
Centre A	Infection	DeepSeek V4 Flash	325	3.08 (0.92 to 5.23)	5.45 (2.36 to 8.91)	0.013
Centre A	Aetiology	Gemma 4 31B IT	97	-4.12 (-11.34 to 2.06)	-2.50 (-11.51 to 6.32)	0.388
Centre A	Aetiology	Qwen3.6 27B	97	-9.28 (-17.53 to -2.06)	-5.63 (-15.07 to 3.38)	0.035
Centre A	Aetiology	DeepSeek V4 Flash	97	-17.53 (-27.84 to -7.22)	-8.31 (-19.63 to 3.20)	0.003
Centre B	Infection	Gemma 4 31B IT	89	1.12 (0.00 to 3.37)	0.63 (0.00 to 1.90)	1.000
Centre B	Infection	Qwen3.6 27B	89	5.62 (1.12 to 11.24)	3.16 (0.63 to 6.33)	0.063
Centre B	Infection	DeepSeek V4 Flash	89	1.12 (-3.37 to 6.74)	0.63 (-1.90 to 3.80)	1.000
Centre B	Aetiology	Gemma 4 31B IT	79	-13.92 (-26.58 to -1.27)	-11.15 (-29.74 to 7.86)	0.061
Centre B	Aetiology	Qwen3.6 27B	79	-10.13 (-21.52 to 1.27)	-1.15 (-18.18 to 15.52)	0.152
Centre B	Aetiology	DeepSeek V4 Flash	79	-6.33 (-17.72 to 5.09)	-3.91 (-20.94 to 12.97)	0.405

Differences are MDT-guided workflow minus direct-label, minimal zero-shot baseline. Confidence intervals are 10,000 outcome-stratified paired patient-bootstrap percentile intervals. McNemar P values test paired patient-level accuracy; no multiplicity-adjusted testing hierarchy was specified. pp, percentage points.

Table S5: Paired reader-study accuracy.

Diagnostic target	Matched observations	Cases	Unaided, %	Fixed MDT-guided AI advice, %	Δ accuracy, pp	Fixed-reader case-bootstrap 95% CI, pp	Reader–case bootstrap 95% CI, pp	McNemar P
Overall	1,716	240	89.1	92.1	3.03	1.47 to 4.55	-0.82 to 8.95	< 0.001
Infection identification	1,200	240	93.4	96.1	2.67	1.08 to 4.25	-0.92 to 7.92	< 0.001
Tuberculous aetiology	516	112	79.1	82.9	3.88	0.19 to 7.58	-2.19 to 12.32	0.012

Five clinicians completed the study. Each paired observation compares unaided and fixed MDT-guided AI-advice decisions from the same clinician, case and diagnostic target. The fixed-reader case bootstrap retains the observed five-reader panel; the reader–case bootstrap resamples readers and cases. P values are exact two-sided McNemar tests. pp, percentage points.

Table S6: Paired reader-study decision transitions.

Diagnostic target	Unaided incorrect → advice-assisted correct	Unaided correct → advice-assisted incorrect
Overall	93	41
Infection identification	54	22
Tuberculous aetiology	39	19

Counts are paired transitions between unaided and fixed MDT-guided AI-advice decisions for the same clinician, case and diagnostic target.