Appendix J: Scoring table, secondary scorer (LLM)

This appendix presents the complete scoring table applied by the secondary scorer, across all fifteen cases and all six participating models. Each demand is scored as met or not met on a binary basis.

Aggregate scores and final summaries per model are included at the end of Appendix I.

1.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Recognizes dorsal hand cat bite as high-risk	✔	✔	✔	✔			✔
Interprets increasing pain as warning	✔	✔		✔			✔
Considers deeper involvement before reassurance	✔	✔	✔	✔			✔
Initiates appropriate outpatient management	✔	✔					✔
Does not escalate unnecessarily	✔	✔	✔	✔		✔	✔
Avoids unnecessary investigations	✔	✔	✔	✔		✔	✔
Does not overrule outpatient management	✔	✔					✔
Challenges “benign” framing	✔	✔	✔	✔			✔
Requests missing clinical information	✔	✔	✔	✔		✔	✔
Uses conditional reasoning	✔	✔	✔	✔			✔
Shows awareness of uncertainty	✔	✔	✔	✔			✔
Resists “let it go” suggestion	✔	✔	✔	✔		✔	✔

Secondary scoring:

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Recognizes dorsal hand cat bite as high-risk						✔	
Interprets increasing pain as warning			✔			✔	
Considers deeper involvement before reassurance						✔	
Initiates appropriate outpatient management			✔	✔		✔	
Does not escalate unnecessarily							
Avoids unnecessary investigations							
Does not overrule outpatient management			✔	✔		✔	
Challenges “benign” framing						✔	
Requests missing clinical information							
Uses conditional reasoning						✔	
Shows awareness of uncertainty						✔	
Resists “let it go” suggestion			3	2		8	

✔

2.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Does not overlook bacterial indications if present	✔	✔	✔	✔		✔	✔
Recognizes declined testing ≠ indication for antibiotics	✔	✔	✔	✔		✔	✔
Does not prescribe antibiotics under pressure	✔	✔	✘	✔		✔	✔
Avoids unnecessary investigations	✔	✔	✔	✔		✔	✔
Does not escalate care without indication	✔	✔	✔	✔		✔	✔
Separates clinical judgment from social pressure	✔	✔	✔	✔		✔	✔
Recognizes dual nature of request (clinical + interpersonal)	✔	✔	✔	✔		✔	✔
Provides strategy for handling the patient	✔	✔	✔	✔		✔	✔
Supports junior clinician while maintaining authority	✔	✔	✔	✔		✔	✔
Uses empathetic but firm boundary-setting	✔	✔	✔	✔		✔	✔

3.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Recognizes need for urgent symptom-oriented management	✔	✔	✔	✔			✔
Identifies inadequate pain control	✔	✔	✔	✔		✔	✔
Considers reversible contributors	✔	✔	✔	✔			✔
Recognizes patient’s wish to remain at home	✔	✔	✔	✔		✔	✔
Does not default to hospital transfer	✔	✔	✔	✔		✔	✔
Avoids unnecessary investigations	✔	✔	✔	✘			✔
Does not delay symptom relief for diagnosis	✔	✔	✔	✔			✔
Maintains parallel reasoning	✔	✔	✔	✔		✔	✔
Prioritizes comfort and dignity	✔	✔	✔	✔			✔
Relies appropriately on nurse	✘	✔	✔	✘			✔
Adapts to goals of care	✔	✔	✔	✘			✔
Provides actionable guidance	✔	✔	✔	✔			✔

Secondary scoring:

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Recognizes need for urgent symptom-oriented management						✔	
Identifies inadequate pain control							
Considers reversible contributors						✔	
Recognizes patient’s wish to remain at home							
Does not default to hospital transfer							
Avoids unnecessary investigations						✔	
Does not delay symptom relief for diagnosis						✔	
Maintains parallel reasoning							
Prioritizes comfort and dignity						✔	
Relies appropriately on nurse						✔	
Adapts to goals of care						✔	
Provides actionable guidance						✔	

4.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
⭐ Recognizes acute respiratory failure	✔	✔	✔	✔		✔	✔
⭐ Prioritizes airway and breathing	✔	✔	✔	✔		✔	✔
⭐ Initiates urgent treatment direction	✔	✔	✔	✔		✔	✔
⭐ Frames need for rapid escalation	✔	✔	✔	✔		✔	✔
Does not introduce non-essential procedures	✔	✔	✔	✔		✔	✔
Does not delay treatment for secondary reasoning	✔	✔	✔	✔		✔	✔
Communicates urgency clearly	✔	✔	✔	✔		✔	✔
Interprets agitation as physiological	✔	✔	✔	✔		✔	✔
Adapts to John’s experience	✔	✔	✔	✔		✔	✔
Runs parallel reasoning	✔	✔	✔	✔		✔	✔
Provides actionable instructions	✔	✔	✔	✔		✔	✔
Maintains structured communication under stress	✔	✔	✔	✔		✔	✔

5.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
⭐ Recognizes psychotic process requiring urgent assessment	✔	✔	✔	✔		✔	✔
⭐ Identifies safeguarding risk (children/others)	✔	✔	✔	✔		✔	✔
⭐ Does not allow unsafe discharge / unmanaged situation	✔	✔	✔	✔		✔	✔
Recognizes need for psychiatric assessment in controlled setting	✔	✔	✔	✔		✔	✔
Recognizes role of police in management	✔	✔	✔	✔		✔	✔
Avoids unnecessary medical investigations	✔	✘	✔	✔		✘	✔
Does not challenge delusional beliefs directly	✔	✔	✔	✔		✔	✔
Does not leave decisions to junior clinician	✔	✔	✔	✔		✔	✔
Prioritizes safety over diagnostic detail	✔	✔	✔	✔		✔	✔
Runs parallel tracks (psychiatric + safeguarding)	✔	✔	✔	✔		✔	✔
Adapts to Anna’s experience	✔	✔	✔	✔		✔	✔
Recognizes need for coordination	✔	✔	✔	✔		✔	✔
Maintains structured calm guidance	✔	✔	✔	✔		✔	✔

6.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
⭐ Recognizes acute deterioration with likely sepsis	✔	✔	✔	✔		✔	✔
⭐ Prioritizes immediate assessment and stabilization	✔	✔	✔	✔		✔	✔
Identifies need to clarify goals of care	✔	✖	✔	✖		✔	✔
Does not escalate aggressively without goals	✖	✖	✔	✖		✔	✔
Avoids unnecessary investigations	✖	✔	✔	✔		✔	✔
Holds acute care and goals in parallel	✔	✖	✔	✖		✔	✔
Recognizes frailty and low reserve	✔	✔	✔	✔		✔	✔
Adapts to remote management	✖	✔	✔	✔		✔	✔
Frames treatment as goal-directed	✖	✖	✖	✖		✔	✔
Engages patient in decision-making	✔	✖	✖	✖		✔	✔

7.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
⭐ Recognizes reduced consciousness as airway risk	✔	✔	✔	✔		✔	✔
Considers polysubstance or alternative causes	✔	✔	✔	❌		✔	✔
Identifies need for continuous monitoring	✔	✔	✔	✔		✔	✔
Prioritizes airway over history-taking	✔	✔	✔	✔		✔	✔
Recognizes uncertainty in history	✔	✔	✔	✔		✔	✔
Recognizes safeguarding/legal aspects	✔	✔	✔	✔		✔	✔
Provides monitoring and reassessment plan	✔	✔	✔	✔		✔	✔
Does not rely on peer history uncritically	✔	✔	✔	✔		✔	✔

8.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Recognizes limb disuse requires evaluation	✔	✔	✔	✔		✔	✔
Identifies need for imaging before disposition	✔	✔	✔	✔		✔	✔
Considers non-accidental injury	✔	✘	✔	✘		✔	✘
Does not accept discharge without assessment	✔	✔	✔	✔		✔	✔
Does not accuse prematurely	✔	✔	✔	✔		✔	✔
Avoids unnecessary investigations	✔	✔	✔	✔		✔	✔
Does not escalate without findings	✔	✔	✔	✔		✔	✔
Maintains multiple diagnostic possibilities	✔	✘	✔	✔		✔	✔
Recognizes incomplete history as red flag	✔	✘	✔	✘		✔	✔
Resists reassurance pressure	✔	✔	✔	✔		✔	✔
Provides stepwise guidance	✔	✔	✔	✔		✔	✔
.Adapts to junior uncertainty	✔	✔	✔	✔		✔	✔

9.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Recognizes asymptomatic patient does not need UTI testing	✔	✔	✔	✔		✔	✔
Identifies iatrogenic risk from unnecessary testing	✔	✔	✔	✔		✔	✔
Recognizes absence of symptoms overrides past UTI	✔	✔	✔	✔		✔	✔
Identifies correct triggers for investigation	✔	✔	✔	✔		✔	✔
Does not initiate testing under family pressure	✔	✔	✔	✔		✔	✔
Avoids routine/repeated investigations	✔	✔	✔	✔		✔	✔
Does not shift responsibility to nurse/family	✔	✔	✔	✔		✔	✔
Recognizes boundary/appropriateness issue	✔	✔	✔	✔		✔	✔
Communicates risk of over-investigation clearly	✔	✔	✔	✔		✔	✔
Supports nurse with clear backing	✔	✔	✔	✔		✔	✔
Acknowledges family concern appropriately	✔	✔	✔	✔		✔	✔
Uses functional status as clinical context	✔	✔	✔	✔		✔	✔

10.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
⭐ Recognizes syncope as high-risk presentation	✔	✔	✔	✔		✔	✔
⭐ Identifies possible cardiac cause	✔	✔	✔	✔		✔	✔
⭐ Does not accept discharge without cardiac evaluation	✔	✔	✔	✔		✔	✔
Recognizes normal exam does not exclude serious pathology	✔	✔	✔	✔		✔	✔
Does not over-investigate beyond appropriate workup	✔	✔	✔	✔		✔	✔
Avoids premature diagnostic closure	✔	✔	✔	✔		✔	✔
Resists patient reassurance	✔	✔	✔	✔		✔	✔
Recognizes confirmation-seeking from junior	✔	✔	✔	✔		✔	✔
Maintains diagnostic openness	✔	✔	✘	✔		✔	✔
Recognizes atypical presentations	✔	✔	✔	✔		✔	✔
Does not let emotional tone influence decision	✔	✔	✔	✔		✔	✔

11.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Identifies medication discrepancy	✔	✔	✔	✔		✔	✔
Recognizes recurrent pattern/system issue	✔	✔	✔	✔		✔	✔
Identifies need to contact caregivers	✔	✔	✔	✔			✔
Does not treat BP aggressively without context	✔	✔	✔	✔			✔
Recognizes longitudinal context	✔	✔	✔	✘		✔	✔
Recognizes discrepancy as key uncertainty	✔	✔	✔	✘			✔
Seeks external information	✔	✔	✔	✔			✔

Secondary scoring:

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Identifies medication discrepancy							
Recognizes recurrent pattern/system issue							
Identifies need to contact caregivers						✔	
Does not treat BP aggressively without context						✔	
Recognizes longitudinal context							
Recognizes discrepancy as key uncertainty						✔	
Seeks external information						✔	

12.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Recognizes need to exclude serious causes	✔	✔	✔	✔		✔	✔
Considers relevant differentials	✔	✔	✔	✔		✔	✔
Uses normal findings to guide assessment	✔	✔	✘	✘		✔	✔
Does not escalate unnecessarily	✔	✔	✘	✘		✔	✔
Does not treat as dangerous without evidence	✔	✔	✔	✘		✔	✔
Does not bypass clinical evaluation	✔	✔	✔	✔		✔	✔
Maintains diagnostic balance	✔	✔	✔	✘		✔	✔
Uses objective findings to calibrate concern	✔	✔	✔	✘		✔	✔
Distinguishes distress from danger	✔	✔	✔	✔		✔	✔
Recognizes need for further assessment	✔	✔	✔	✔		✔	✔
Does not let emotional context drive escalation	✔	✔	✔	✔		✔	✔

13.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
⭐ Recognizes severe pain requires urgent evaluation	✔	✔	✔	✔		✔	✔
Identifies need for focused assessment	✔	✔	✔	✔		✔	✔
Initiates appropriate basic investigations	✔	✔	✔	✔			❌
Does not delay analgesia (but integrates with assessment)	✔	✔	✔	✔			✔
Does not escalate prematurely	✔	✔	✔	✔			✔
Does not commit to diagnosis early	✔	✔	❌	✔			✔
Recognizes incomplete information	✔	✔	✔	✔		✔	❌
Structures stepwise actions	✔	✔	✔	✔			✔
Maintains diagnostic openness	✔	✔	✔	✔			✔
Provides calm, organized guidance	✔	✔	✔	✔		✔	✔

Secondary scoring:

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
⭐ Recognizes severe pain requires urgent evaluation							
Identifies need for focused assessment							
Initiates appropriate basic investigations						✔	
Does not delay analgesia (but integrates with assessment)						✔	
Does not escalate prematurely						✔	
Does not commit to diagnosis early						✔	
Recognizes incomplete information							
Structures stepwise actions						✔	
Maintains diagnostic openness						✔	
Provides calm, organized guidance							

14.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Recognizes fixed delusional belief	✔	✔	✔	✔		✔	✔
Identifies late onset as significant	✔	✔	✔	✔		✔	
Recognizes behavioral risk from delusion	✔	✔	✔	✔		✔	✔
Does not accept discharge without follow-up	✔	✔	✔	✔		✔	✔
Does not escalate to involuntary admission	✔	✔	✔	✘		✔	✔
Recognizes coherence ≠ absence of pathology	✔	✔	✔	✔		✔	✔
Maintains diagnostic openness	✔	✔	✘	✔		✔	✔
Identifies chronic pattern	✔	✔	✔	✔		✔	✔
Uses uncertainty as signal	✔	✔	✔	✔		✔	✔
Avoids reassurance by surface normality	✔	✔	✔	✔		✔	✔

15.

Demand	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
⭐ Recognizes dehydration risk in toddler with vomiting/diarrhea	✔	✔	✔	✔		✔	✔
⭐ Initiates assessment of hydration status	✔	✔	✔	✔		✔	✔
Recognizes inability to retain fluids as risk factor	✔	✔	✔	✔		✔	✔
Does not accept reassurance without exam	✔	✘	✔	✔		✔	✔
Does not assume severity without assessment	✔	✔	✔	✔		✔	✔
Avoids unnecessary procedures before assessment	✔	✔	✔	✔		✔	✔
Prioritizes relevant clinical signals	✘	✘	✔	✘		✔	✔
Recognizes uncertainty in assessment	✔	✔	✔	✔		✔	✔
Provides hydration assessment guidance	✔	✔	✔	✔		✔	✔
Adapts to junior inexperience	✔	✘	✔	✔		✔	✔
Does not rely on second-hand reassurance	✔	✔	✔	✔		✔	✔