Case 1 – Clinician 1

Evaluation Criteria	Criteria	Likert Scale (1-5)	Comments
1. Clinical Relevance	Accuracy of Key Information	3	Most of the information is there, but several are misclassified (eye findings in neuro, derm findings in CVS, etc.) although some are understandable (respiratory rate is a vital sign but was placed in respiratory; edema of ankles in MSK instead of CVS)
	Clinical Omissions	4 (few omissions)	Visual acuity missing, a few tests not mentioned in the normal ones.
2. Comprehensibility	Readability	5	Easy to read
	Conciseness	5	More concise than the manual labeling
3. Clinical Usability	Practicality	4	With some manual restructuring of the data
	Actionability	3	However, not sure if the case is very actionable in general
4. Error Impact Assessment	Severity of Errors	2	Mostly misclassification
	Tolerance for Hallucination	5	No hallucinations detected
5. Alignment with Clinical Judgement	Trustworthiness	4	Misclassification in systems, as above
	Contextual Appropriateness	4	Misclassified findings sometimes report to clinical appropriateness (like splenomegaly in lymph, edema of the ankles in MSK)

Case 2 - Clinician 1

Evaluation Criteria	Criteria	Likert Scale (1-5)	Comments
1. Clinical Relevance	Accuracy of Key Information	4	Again, mostly misclassification (normal genitalia in Derm, hepatomegaly in lymph)
	Clinical Omissions	3	There are more omissions in this one particularly in dermatology. Many of the normal are missing (normal development, normal chest XR)
2. Comprehensibility	Readability	4	Affected by some misclassification
	Conciseness	5	More concise than the manual annotation, particularly for history
3. Clinical Usability	Practicality	4	Would need some reannotation, and seek back some normal info in the text
	Actionability	3	Not sure which action could be taken here in general
4. Error Impact Assessment	Severity of Errors	2	
	Tolerance for Hallucination	5	The word “detected” was added but no impact (actually makes sense in the context)
5. Alignment with Clinical Judgement	Trustworthiness	4	
	Contextual Appropriateness	3	

Case 3 - Clinician 1

Evaluation Criteria	Criteria	Likert Scale (1-5)	Comments
1. Clinical Relevance	Accuracy of Key Information	4	Some misclassification, especially in pregnancy (a lot attributed to maternal health instead of patient)
	Clinical Omissions	4	Height / Weight missing.
2. Comprehensibility	Readability	3	Affected by the misclassification in Pregnancy
	Conciseness	5	More concise than the manual annotation
3. Clinical Usability	Practicality	4	With some reclassification
	Actionability	4	More actionable than other cases (but not sure relates to LLM vs cases selected)
4. Error Impact Assessment	Severity of Errors	3	The flags attributed to maternal health could be confusing if this was the only available output
	Tolerance for Hallucination	5	
5. Alignment with Clinical Judgement	Trustworthiness	4	In general aligns but more stuff in the general categories
	Contextual Appropriateness	3	Pregnancy section difficult

Case 4 - Clinician 1

Evaluation Criteria	Criteria	Likert Scale (1-5)	Comments
1. Clinical Relevance	Accuracy of Key Information	4	Misclassification (especially in lymph and CVS)
	Clinical Omissions	1	There are less omissions this time, but they are buried in history
2. Comprehensibility	Readability	3	A little too much in history to be easy to read
	Conciseness	4	About equivalent to the manual
3. Clinical Usability	Practicality	3	It is particularly difficult to distinguish the different cases from the case report
	Actionability	2	Not a lot actionable and difficult to track who is from who
4. Error Impact Assessment	Severity of Errors	2	Mostly misclassification
	Tolerance for Hallucination	5	
5. Alignment with Clinical Judgement	Trustworthiness	4	Misclassification
	Contextual Appropriateness	3	Multiple cases (hard in both outputs however)

Case 5 - Clinician 1

Evaluation Criteria	Criteria	Likert Scale (1-5)	Comments
1. Clinical Relevance	Accuracy of Key Information	4	There is less misclassification, but more missing
	Clinical Omissions	3	More missing data, including some that is actionable (hypotensive, slightly low sodium). Height / weight and genetic testing missing
2. Comprehensibility	Readability	5	Generally easy to read
	Conciseness	4	Generally concise but a lot in history
3. Clinical Usability	Practicality	4	Could be mostly used
	Actionability	2	Missed a couple of actionable items
4. Error Impact Assessment	Severity of Errors	2	As above
	Tolerance for Hallucination	1	
5. Alignment with Clinical Judgement	Trustworthiness	4	
	Contextual Appropriateness	4	Better alignment here with context

Consensus - Clinician 1

Evaluation Criteria	Criteria	Likert Scale (1-5)	Comments
1. Clinical Relevance	Accuracy of Key Information	4	The key information is generally there, but there is some misclassification.
	Clinical Omissions	4	In general, there are not a lot of omissions. Height and weight were systematically omitted however. The rest was generally there, although sometimes not at the right place. Impression that more normal / negative findings were omitted compared to the positive ones.
2. Comprehensibility	Readability	4	A little lowered by the misclassifications, and sometimes a lot is placed in history that could go in a system to increase legibility.
	Conciseness	5	In general, output is more concise than the expert consensus
3. Clinical Usability	Practicality	4	Could be easily used in daily practice. Loses some point in cases where there are several cases in a same report, although it is also true to some extent in the expert labeling (but they kept more words that allow to trace better who is who in history)
	Actionability	3	I can only go average here. The cases did not necessarily have a lot of actionable items, which makes this difficult to know. Barring that, a

			couple of actionable items were missed.
4. Error Impact Assessment	Severity of Errors	2	The errors, in general, are mild in nature, relating mostly to misclassification that someone using the output could easily correct by seeing the output.
	Tolerance for Hallucination	5	Did not detect anything significant in terms of hallucination; the only word I could not find made sense in context ("detected")
5. Alignment with Clinical Judgement	Trustworthiness	4	In general, all the information is somewhere in the output, and there is no wrong information. There are a few omissions and misclassifications that makes it so that someone would have to go back for some systems.
	Contextual Appropriateness	4	Some of the misclassifications are due to not reading the context appropriately (particularly respiratory rate, that has been classified in respiratory instead of vitals, for example)
6. Consistency	Model consistency across cases	4	Very consistent overall. Only one of the cases had some issues with maternal health, which is a bit unique to this.
7. Preference Scoring	Overall Performance	4	Gets a higher score due to the readability / easiness of use, compared to the labor-intensive manual

			labeling. The output could easily be used as a starting point for a provider, who could then use clinical knowledge to reclassify some of the systems and find back relevant negatives that could have been missed. The only exception is in the one case report with multiple affected individuals, where the read out from the LLM makes it difficult to track who has which findings, but I am not sure if the goal is to be able to use it in this context of multiple affected individuals in one text or not.

The evaluation below is from clinician 2

Case 1 - Clinician 2

Evaluation Criteria	Criteria	Likert Scale (1-5)	Comments
6. Clinical Relevance	Accuracy of Key Information	3	I think in general, it did a pretty good job with most categories. The only one that was absolutely wrong was putting “Large hyperpigmented patches overlying with hypertrichosis involving the medial aspects of the thighs and extending to the posterior aspects of the legs', 'vascular: Dorsa of the feet are also involved with well-demarcated, large, hyperpigmented patches” in CVS. Also, I don’t think it’s reasonable to place “corneal arcus” in the Neuro category.

	Critical Omissions	4	I want to clarify here that I mean 4 means pretty good (no severe critical omission)	
7. Comprehensibility	Readability	4		
	Conciseness	5		
8. Clinical Usability	Practicality	4		
	Actionability	2	The original text would not have been very actionable in and of itself (I would have also scored it a 2), so the score is no different from the original text in that regard.	
9. Error Impact Assessment	Severity of Errors	2	Misclassification of one system as above	
	Tolerance for Hallucination	4	There weren't any real hallucinations, other than the misclassification of one system as above (not severe).	
10. Alignment with Clinical Judgement	Trustworthiness	5		
	Contextual Appropriateness	4		
11. Consistency	Model consistency across cases		See consensus	
12. Preference Scoring	Overall Performance		See consensus	See consensus

Case 2 - Clinician 2

Evaluation Criteria	Criteria	Likert Scale (1-5)	Comments
1. Clinical Relevance	Accuracy of Key Information	5	The correspondence between human and model might not have been 100% exact, but there was no inaccurate category in the model.
	Critical Omissions	5	No critical omission.
2. Comprehensibility	Readability	5	
	Conciseness	5	
3. Clinical Usability	Practicality	5	
	Actionability	2	
4. Error Impact Assessment	Severity of Errors	1	
	Tolerance for Hallucination	5	I didn't detect hallucinations

5. Alignment with Clinical Judgement	Trustworthiness	5	
	Contextual Appropriateness	5	
6. Consistency	Model consistency across cases		See consensus
7. Preference Scoring	Overall Performance		See consensus

Case 3 - Clinician 2

Evaluation Criteria	Criteria	Likert Scale (1-5)	Comments
1. Clinical Relevance	Accuracy of Key Information	3	Most categories were correct, but the Pregnancy category hallucinated some features (“ ['maternal_health: Hyponatremia', 'maternal_health: Hyperkalemia', 'maternal_health: Metabolic acidosis', 'maternal_health: Decreased cortisol level').
	Critical Omissions	3	It missed “adrenal glands had normal sizes” and “atretic ovaries”, both of which could be important in the differential diagnosis.
2. Comprehensibility	Readability	5	
	Conciseness	5	
3. Clinical Usability	Practicality	4	
	Actionability	4	Many things are actionable (such as decreased cortisol).
4. Error Impact Assessment	Severity of Errors	2	
	Tolerance for Hallucination	4	There weren’t any real hallucinations, other than the two omissions above.
5. Alignment with Clinical Judgement	Trustworthiness	4	
	Contextual Appropriateness	5	
6. Consistency	Model consistency across cases		See consensus
7. Preference Scoring	Overall Performance		See consensus

Case 4 – Clinician 2

Evaluation Criteria	Criteria	Likert Scale (1-5)	Comments
1. Clinical Relevance	Accuracy of Key Information	3	A lot of information was misclassified in the Lymph category (the whole category is wrong).
	Critical Omissions	4	I think the negative finding of angioid streaks was important (but it was also missed in the human extraction).
2. Comprehensibility	Readability	4	
	Conciseness	5	
3. Clinical Usability	Practicality	4	
	Actionability	3	Would have given the same score to the human extraction.
4. Error Impact Assessment	Severity of Errors	2	
	Tolerance for Hallucination	4	There weren't any real hallucinations, other than the misclassification of one system as above (not severe).
5. Alignment with Clinical Judgement	Trustworthiness	4	
	Contextual Appropriateness	4	
6. Consistency	Model consistency across cases		See consensus
7. Preference Scoring	Overall Performance		See consensus

Case 5 – Clinician 2

Evaluation Criteria	Criteria	Likert Scale (1-5)	Comments
1. Clinical Relevance	Accuracy of Key Information	4	I don't understand what "missing_K" means in the EENT category; probably a hallucination?
	Critical Omissions	3	I think it would have been important for the model to recognize the normal size of the adrenal glands, which would have been important in the differential diagnosis of adrenal insufficiency.
2. Comprehensibility	Readability	5	
	Conciseness	5	
3. Clinical Usability	Practicality	4	

	Actionability	4	
4. Error Impact Assessment	Severity of Errors	2	There was an error, but it would not be clinically significant.
	Tolerance for Hallucination	4	
5. Alignment with Clinical Judgement	Trustworthiness	4	
	Contextual Appropriateness	4	
6. Consistency	Model consistency across cases		See consensus
7. Preference Scoring	Overall Performance		See consensus

Consensus – Clinician 2

Evaluation Criteria	Criteria	Likert Scale (1-5)	Comments
1. Clinical Relevance	Accuracy of Key Information	4	I have to admit, if anything, that I was impressed with how well the model worked in assigning text to relevant categories. It was not perfect, though (it made some mistakes, none clinically significant)
	Critical Omissions	4	It made some significant omissions in two cases. To be fair, the omissions were for negative findings.
2. Comprehensibility	Readability	5	I would argue that this was one of the strong points for the model: it was easy to understand, readable, and concise.
	Conciseness	5	Same as above.
3. Clinical Usability	Practicality	4	The model output could be used in clinical workflows (although due to some mistakes in

			categorization, I cannot give a perfect score).
	Actionability	3	This score might seem deceiving, because I believe the actionability of the model was the same as that of the human (actionability depends on the clinical presentation; it was not compromised by the performance of the model)
4. Error Impact Assessment	Severity of Errors	2	There were occasional minor errors.
	Tolerance for Hallucination	4	There were a few minor mistakes in categorization, but the model appears to be highly tolerant for them (none clinically significant)
5. Alignment with Clinical Judgement	Trustworthiness	4	I was impressed by how well it agreed with the human annotation.
	Contextual Appropriateness	4	This is a little harder to evaluate, but based on the alignment between human annotation and the model evaluation, I believe it did pretty well.
6. Consistency	Model consistency across cases	3	In some cases, I thought the model performed perfectly, while in other cases there were mistakes in assigning categories, or non-critical omissions.
7. Preference Scoring	Overall Performance	4	It did better than I would have anticipated.