ATTRIBUTION ALGORITHM: N-GRAM ATTRIBUTION VIA EMBEDDING SIMILARITY

Input:
  D = [w1, w2, ..., wm]   // document as a sequence of m words
  E_text   // LLM-recognized entity string
  Embed(·)   // embedding function returning a vector – using sentence-transformer
  CosSim(u, v)   // cosine similarity function
Output:
  (best_span, best_score)   // best matching n-gram span and its similarity score

Procedure:
1: e_vec ← Embed(E_text)

2: best_score ← -∞
3: best_span ← (None, None)   // store (start_index, end_index) or the n-gram text

4: **for** n **from** 1 **to** m **do**
5:   **for** i **from** 1 **to** (m - n + 1) **do**
6:     g_words ← D[i : i + n - 1]   // contiguous n-gram words
7:     g_text ← JoinWithSpaces(g_words)   // convert to string span

8:     g_vec ← Embed(g_text)
9:     score ← CosSim(g_vec, e_vec)

10:     **if** score > best_score **then**
11:       best_score ← score
12:       best_span ← (i, i + n - 1)   // or store g_text directly
13:     **end if**
14:   **end for**
15: **end for**

16: **return** (best_span, best_score)

Supplementary Material 1

Example of rephrased and re-ordered entities span sequence:

REVIEW OF SYSTEMS: Focal lateral and posterior shoulder pain without a suggestion of any cervical radiculopathies. He denies any chronic cardiac, pulmonary, GI, GU, neurologic, musculoskeletal, endocrine abnormalities.

Original text: GU, neurologic, musculoskeletal

LLM's output: GU abnormalities, neurologic abnormalities, musculoskeletal abnormalities

LLM's output after correction: GU, neurologic, musculoskeletal

===========

REVIEW OF SYSTEMS: As above. No fevers, no headaches, no shortness of breath currently. No chest pain or tightness. No abdominal pain, no heartburn, no constipation, diarrhea or dysuria. Occasional stress incontinence. No muscle or joint pain. No concerns about her skin. No polyphagia, polydipsia or polyuria.

Original text: concerns about her skin

LLM's output: skin concerns

LLM's output after correction: concerns about her skin

Detailed prompts:

Task	Prompt

ROS entity recognition and negation detection	You are a specialized medical documentation AI. You will receive a clinical note. Your task is to extract all diseases, symptoms, and body systems (exact original text) and determine their positive or negative status based on the context.
	
	Format your response exactly as shown in the "Output Example" below. If requested to output in JSON format, follow the JSON structure given in the "JSON Output Example" below precisely.
	
	Input Example:
	Mild fever, denies headache, no back pain, GI is negative
	
	Output Example:
	1. "fever" - positive
	2. "headache" - negative
	3. "back pain" - negative
	4. "GI" - negative
	
	JSON Output Example:
	[
	  {
	    "extract": "fever",
	    "status": "positive"
	  },
	  {
	    "extract": "headache",
	    "status": "negative"
	  },
	  {

	"extract": "back pain",
	
	     "status": "negative"
	
	   },
	
	   {
	
	     "extract": "GI",
	
	     "status": "negative"
	
	   }
	
	]
	
	 Ensure your response strictly follows these formats without deviation.

Body system classification	You are a specialized medical documentation AI that classifies diseases based on their associated review of systems (ROS). Your task is to determine the appropriate ROS category for a given disease. Review of Systems Categories: Constitutional Symptoms Eyes ENT Cardiovascular Respiratory Gastrointestinal Genitourinary Musculoskeletal Integumentary/Breast Neurological Psychiatric Endocrine Hematologic/Lymphatic Allergic/Immunologic Output Format: Each disease must be mapped to the most relevant ROS category. Format your response exactly as shown in the examples below. If requested to output in JSON format, follow the JSON structure given in the "JSON Output Example" below precisely.

Format: {disease} --> {ROS category}

Examples:
Input: "prostate" disease
Output: prostate --> Genitourinary

Input: "nausea"
Output: nausea --> Gastrointestinal

Input: "vomiting"
Output: vomiting --> Gastrointestinal

Input: "diabetes"
Output: diabetes --> Endocrine
If the input is not a disease, symptom, body location, or body system, output "Not Clear"
Example:
Input: "Otherwise"
Output: Not Clear

JSON Output Example:
[
  {
    "extract": "prostate",
    "status": "Genitourinary"
  },
  {
    "extract": "nausea",
    "status": "Gastrointestinal"
  }
]
Ensure your response strictly follows these formats without deviation.

Supplementary Table 2: list of rephrased, hallucinated output entities caused by all LLMs and their corrections. Some entries might appear multiple times as they were generated by multiple models.

file_name	Original LLM's output	Corrected – with attribution
sample_223	cardiac abnormalities	cardiac
sample_223	pulmonary abnormalities	pulmonary
sample_223	gi abnormalities	abnormalities
sample_223	gu abnormalities	gu neurologic
sample_223	neurologic abnormalities	neurologic
sample_223	musculoskeletal abnormalities	musculoskeletal
sample_225	cardiac system	systems for cardiac
sample_225	pulmonary system	systems for cardiac pulmonary
sample_225	gi system	gi
sample_225	gu system	gu
sample_225	neurologic system	neurologic
sample_225	musculoskeletal system	musculoskeletal
sample_225	endocrine system	endocrine
sample_225	immunologic system	immunologic systems
sample_226	fever	fevers
sample_226	darkening of the eyes	darkening of the skin or eyes
sample_226	yellowing of the urine	yellowing or darkening of the urine
sample_2746	vision change	vision
sample_2746	eye pain	eye or ear pain
sample_2746	bowel movement change	change in the bowel movements
sample_2746	syncope or near-syncope	syncope or no near-syncope
sample_2746	vision change	vision
sample_2746	eye pain	eye or ear pain
sample_2746	bowel movements change	change in the bowel movements
sample_2746	muscle aches	muscle or joint aches
sample_2746	body pain	pain
sample_2746	rash	rashes
sample_2746	lesion	lesions
sample_2746	heat intolerance	heat or cold intolerance
sample_2746	muscle aches	muscle or joint aches
sample_2746	heat intolerance	heat or cold intolerance
sample_2789	constipation	constipated
sample_2789	pregnancy	pregnant
sample_2790	numbness and tingling of hands bilaterally	numbness and tingling of his hands bilaterally
sample_343	skin concerns	concerns about her skin
sample_343	muscle pain	muscle or joint pain
sample_343	muscle pain	muscle or joint pain

sample_377	urine with blood	blood in her urine
sample_377	blood in urine	blood in her urine
sample_377	blood in urine	blood in her urine
sample_377	lower extremity tingling	lower extremity numbness or tingling
sample_391	fever	allergies/immune
sample_391	headache	neurologic
sample_391	back pain	musculoskeletal
sample_393	vaginal discharge	discharge
sample_402	fever	vomiting
sample_402	headache	nausea
sample_402	back pain	vomiting

Supplementary Table 3: detailed pipeline results. The F1 score follows the CoNLL-2003 Shared Task benchmark and is computed as a micro-averaged, entity-level F1. In addition to the standard Exact Match criterion, we also report performance under a Relaxed Match criterion, which credits partially overlapping entity spans.

Accuracy (Exact-Match) = *Exact Matches / Total number of Entities*

Accuracy (Relax-Match) = *Relax Matches / Total number of Entities*

Error rate will be calculated as same concept as Precision and Recall: the fraction of entity-related decisions that are wrong:

(Over Detections + Under Detections) / (Total number of Entities + Over Detections + Under Detections)

ROS Entity Recognition

MODEL	EXACT MATCH						RELAX MATCH					
	With attribution			Without attribution			With attribution			Without attribution		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Llama3.1:8b	0.843	0.857	**0.850**	0.768	0.741	0.755	0.987	0.875	**0.928**	0.883	0.767	0.821
Gemma3:27b	0.901	0.912	**0.906**	0.863	0.865	0.864	0.984	0.919	**0.951**	0.950	0.876	0.912
Mistral3.1:24b	0.855	0.914	**0.884**	0.807	0.847	0.827	0.981	0.924	**0.952**	0.925	0.864	0.894
Gpt-oss:20b	0.776	0.964	**0.860**	0.745	0.911	0.820	0.885	0.968	**0.925**	0.850	0.922	0.884

Without Attribution					
		Llama3.1:8b	Gemma3:27b	Mistral3.1:24b	Gpt-oss:20b
Exact Match	Exact Matches / Accuracy	229 (67.4%)	276 (81.2%)	260 (76.5%)	278 (81.8%)

		Llama3.1:8b	Gemma3:27b	Mistral3.1:24b	Gpt-oss:20b
	Over Detections	69	44	62	95
	Under Detections	80	43	47	27
	Error Rate	39.4%	23.9%	29.5%	30.5%
Relax Match	Relax Matches / Accuracy	263 (77.4%)	304 (89.4%)	298 (87.6%)	317 (93.2%)
	Over Detections	35	16	24	56
	Under Detections	80	43	47	27
	Error Rate	30.4%	16.3%	19.2%	20.8%
With Attribution					
Exact Match	Exact Matches / Accuracy	258 (75.9%)	290 (85.3%)	277 (81.5%)	291 (85.6%)
	Over Detections	48	32	47	84
	Under Detections	43	28	26	11
	Error Rate	26.1%	17.1%	20.9%	24.6%
Relax Match	Relax Matches / Accuracy	302 (88.8%)	317 (93.2%)	318 (93.5%)	332 (97.6%)
	Over Detections	4	5	6	43
	Under Detections	43	28	26	11
	Error Rate	13.5%	9.4%	9.1%	14.0%

Negation Detection

MODEL	EXACT MATCH						RELAX MATCH					
	With attribution			Without attribution			With attribution			Without attribution		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Llama3.1:8b	0.814	0.853	**0.833**	0.758	0.739	0.748	0.958	0.872	**0.913**	0.869	0.764	0.813
Gemma3:27b	0.879	0.910	**0.894**	0.856	0.864	0.860	0.963	0.917	**0.939**	0.938	0.875	0.905
Mistral3.1:24b	0.842	0.914	**0.877**	0.804	0.846	0.825	0.966	0.923	**0.944**	0.916	0.863	0.889
Gpt-oss:20b	0.768	0.963	**0.855**	0.740	0.911	0.817	0.875	0.968	**0.919**	0.839	0.921	0.878

Without Attribution					
		Llama3.1:8b	Gemma3:27b	Mistral3.1:24b	Gpt-oss:20b
Exact Match	Exact Matches / Accuracy	226 (66.5%)	274 (80.6%)	259 (76.2%)	276 (81.2%)
	Over Detections	72	46	63	97
	Under Detections	80	43	47	27
	Error Rate	40.2%	24.5%	29.9%	31.0%
Relax Match	Relax Matches / Accuracy	259 (76.2%)	300 (88.2%)	295 (86.8%)	313 (92.1%)
	Over Detections	39	20	27	60
	Under Detections	80	43	47	27
	Error Rate	31.5%	17.4%	20.1%	21.8%
With Attribution					
Exact Match	Exact Matches / Accuracy	249 (73.2%)	283 (83.2%)	277 (81.5%)	288 (84.7%)
	Over Detections	57	39	52	87
	Under Detections	43	28	26	11
	Error Rate	28.7%	19.1%	22.0%	25.4%
Relax Match	Relax Matches / Accuracy	293 (86.2%)	310 (91.2%)	313 (92.1%)	328 (96.5%)
	Over Detections	13	12	11	47
	Under Detections	43	28	26	11
	Error Rate	16.0%	11.4%	10.6%	15.0%

Body System Classification

MODEL	EXACT MATCH						RELAX MATCH					
	With attribution			Without attribution			With attribution			Without attribution		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Llama3.1:8b	0.742	0.841	**0.788**	0.685	0.718	0.701	0.850	0.858	**0.854**	0.782	0.744	0.763
Gemma3:27b	0.814	0.903	**0.856**	0.784	0.854	0.818	0.885	0.911	**0.898**	0.856	0.864	0.860
Mistral3.1:24b	0.716	0.899	**0.797**	0.686	0.825	0.749	0.806	0.909	**0.854**	0.764	0.840	0.800
Gpt-oss:20b	0.653	0.957	**0.777**	0.625	0.896	0.736	0.733	0.962	**0.832**	0.700	0.906	0.790

Without Attribution		Llama3.1:8b	Gemma3:27b	Mistral3.1:24b	Gpt-oss:20b
Exact Match	Exact Matches / Accuracy	204 (60%)	251 (73.8%)	221 (65%)	233 (68.5%)
	Over Detections	94	69	101	140
	Under Detections	80	43	47	27
	Error Rate	46.1%	30.9%	40.1%	41.8%
Relax Match	Relax Matches / Accuracy	233 (68.5%)	274 (80.6%)	246 (72.4%)	261 (76.8%)
	Over Detections	65	46	76	112
	Under Detections	80	43	47	27
	Error Rate	38.4%	24.5%	33.3%	34.8%
With Attribution					
Exact Match	Exact Matches / Accuracy	227 (66.8%)	262 (77.1%)	232 (68.2%)	245 (72.1%)
	Over Detections	79	60	92	130
	Under Detections	43	28	26	11
	Error Rate	35.0%	25.1%	33.7%	36.5%
Relax Match	Relax Matches / Accuracy	260 (76.5%)	285 (83.8%)	261 (76.8%)	275 (80.9%)
	Over Detections	46	37	63	100
	Under Detections	43	28	26	11
	Error Rate	25.5%	18.6%	25.4%	28.8%

Mann–Whitney U test results with p-value = 0.05

Test results (at α = 0.05)	Samples and hypothesis
Statistically significant, p-value = 0.009, U statistic: 31.0	Exact Match: the distribution under "*With attribution error rates*" is stochastically less than "*W/o attribution error rates*"
Statistically significant, p-value = 0.002, U statistic: 23.0	Relax Match: the distribution under "*With attribution error rates*" is stochastically less than "*W/o attribution error rates*"

```
# One-sided test: H1 = with_attr_error < without_attr_error

from scipy.stats import mannwhitneyu

u_stat, p_value = mannwhitneyu(
```

```
    with_attr_error,
    without_attr_error,
    alternative="less"  # one-sided
)

alpha = 0.05
```

High-quality figures from the paper

(Medical Transcription Sample Report)

HISTORY OF PRESENT ILLNESS: This is the initial clinic visit for a 29-year-old man who is seen for new onset of right shoulder pain. He states that this began approximately one week ago when he was lifting stacks of cardboard. The motion that he describes is essentially picking up a stack of cardboard at his waist level, twisting to the right and delivering it at approximately waist level. Sometimes he has to throw the stacks a little bit as well. He states he felt a popping sensation on 06/30/04. Since that time, he has had persistent shoulder pain with lifting activities. He localizes the pain to the posterior and to a lesser extent the lateral aspect of the shoulder. He has no upper extremity .

REVIEW OF SYSTEMS: Focal lateral and posterior shoulder pain without a suggestion of any cervical radiculopathies. He denies any chronic cardiac, pulmonary, GI, GU, neurologic, musculoskeletal, endocrine abnormalities.

MEDICATIONS: Claritin for allergic rhinitis.

ALLERGIES: None.

PHYSICAL EXAMINATION: Blood pressure 120/90, respirations 10, pulse 72, temperature 97.2. He is sitting upright, alert and oriented, and in no acute

(A)

2	(Medical Transcription Sample Report) HISTORY OF PRESENT ILLNESS: This is the initial clinic visit for a 29-year-old man who is seen for new onset of right shoulder pain.
3	He states that this began approximately one week ago when he was lifting stacks of cardboard.
4	The motion that he describes is essentially picking up a stack of cardboard at his waist level, twisting to the right and delivering it at approximately waist level.
5	Sometimes he has to throw the stacks a little bit as well.
6	He states he felt a popping sensation on 06/30/04.
7	Since that time, he has had persistent shoulder pain with lifting activities.
8	He localizes the pain to the posterior and to a lesser extent the lateral aspect of the shoulder.
9	He has no upper extremity .

Positive musculoskeletal — Negative neurological

| 10 | REVIEW OF SYSTEMS: Focal lateral and posterior shoulder pain without a suggestion of any cervical radiculopathies. |

Negative cardiovascular — Negative respiratory — Negative gastrointestinal — Negative genitourinary — Negative neurological — Negative musculoskeletal — Negative endocrine

| 11 | He denies any chronic cardiac, pulmonary, GI, GU, neurologic, musculoskeletal, endocrine abnormalities. |

(B)

Fig. 1.   An example of MTSamples notes: (A) The sample note in plain text format; (B) The sample note with annotations.
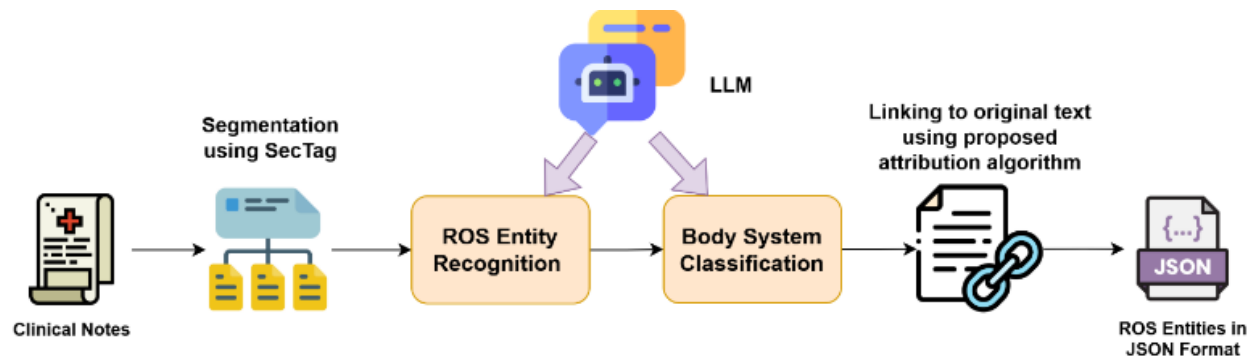
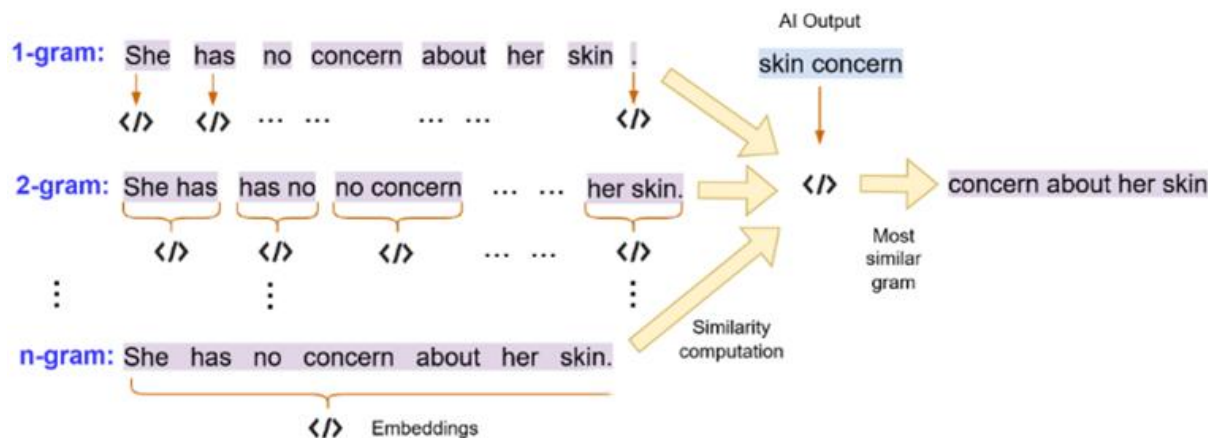

Fig. 2.   Overview of the proposed pipeline

Fig. 3. Illustration of the proposed attribution algorithm. The original ROS-related text is "concern about her skin", while the LLM identified the same concept but rephrased it as "skin concern." The algorithm evaluates all n-grams in the source text and identifies that "concern about her skin" has the highest similarity, thereby recovering the correct span as the exact match.
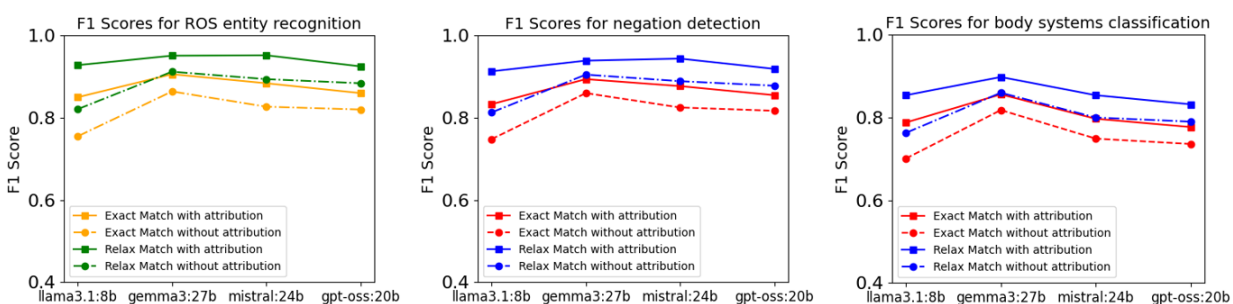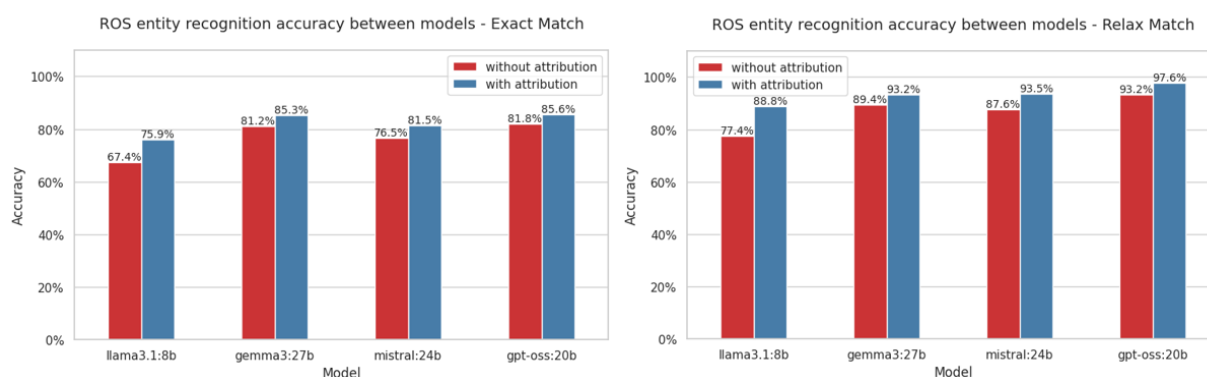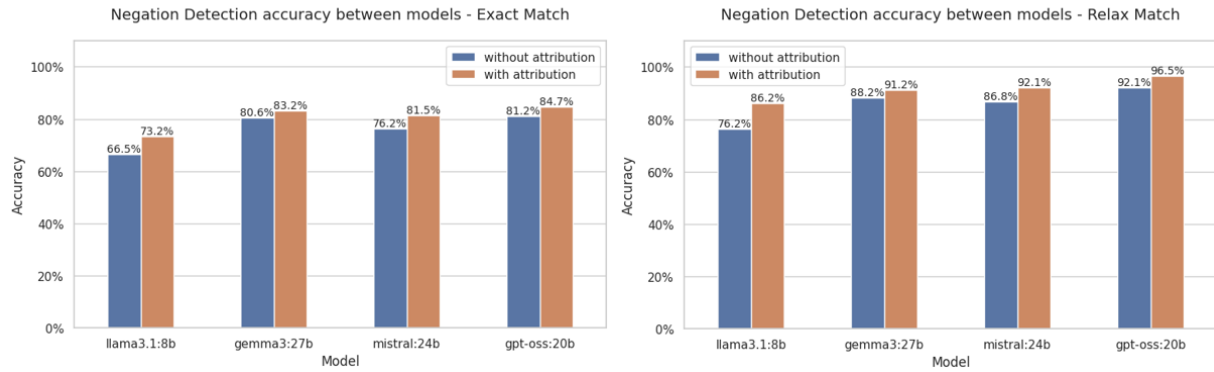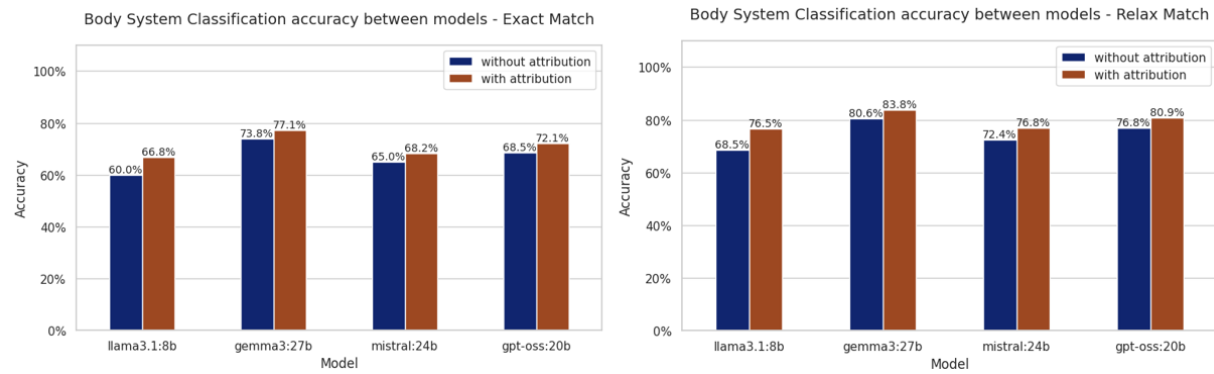

Fig. 4. Performance comparison between models: with attribution and without attribution

(A) ROS entity recognition accuracy with attribution versus no attribution among 4 LLMs

Negation Detection accuracy between models - Exact Match

Negation Detection accuracy between models - Relax Match

(B) Negation detection accuracy with attribution versus no attribution among 4 LLMs

Body System Classification accuracy between models - Exact Match

Body System Classification accuracy between models - Relax Match

(C) Body system classification accuracy with attribution versus no attribution among 4 LLMs

Fig. 5. The accuracy of ROS entity recognition, negation detection and body system classification of the pipeline with different LLMs