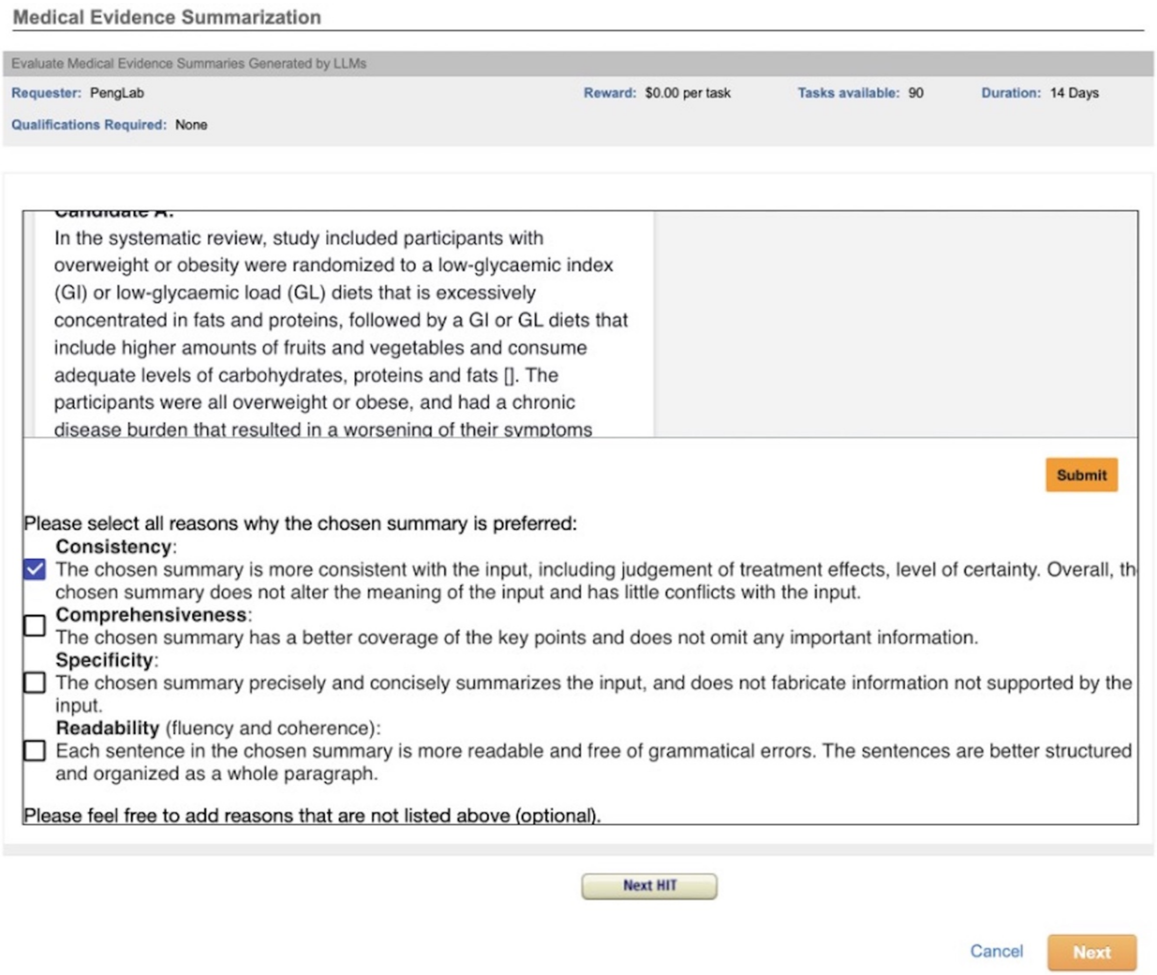
SUPPLEMENTAL MATERIALS



Medical Evidence Summarization

Evaluate Medical Evidence Summaries Generated by LLMs

Requester: PengLab

Reward: $0.00 per task

Tasks available: 90

Duration: 14 Days

Qualifications Required: None

Candidate A:

In the systematic review, study included participants with overweight or obesity were randomized to a low-glycaemic index (GI) or low-glycaemic load (GL) diets that is excessively concentrated in fats and proteins, followed by a GI or GL diets that include higher amounts of fruits and vegetables and consume adequate levels of carbohydrates, proteins and fats []. The participants were all overweight or obese, and had a chronic disease burden that resulted in a worsening of their symptoms

Please select all reasons why the chosen summary is preferred:

☑ **Consistency**: The chosen summary is more consistent with the input, including judgement of treatment effects, level of certainty. Overall, th chosen summary does not alter the meaning of the input and has little conflicts with the input.

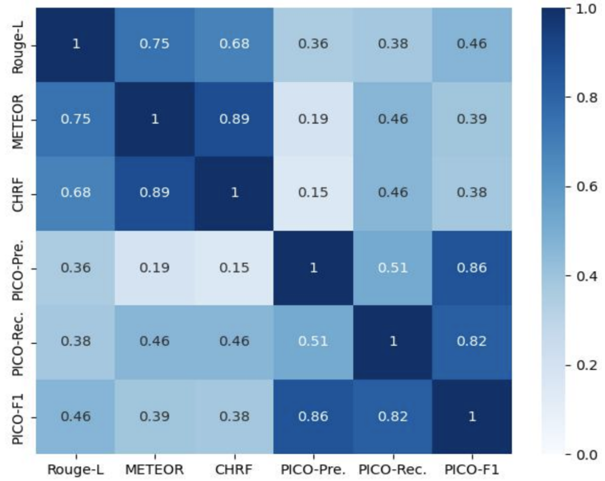
□ **Comprehensiveness**: The chosen summary has a better coverage of the key points and does not omit any important information.

□ **Specificity**: The chosen summary precisely and concisely summarizes the input, and does not fabricate information not supported by the input.

□ **Readability** (fluency and coherence): Each sentence in the chosen summary is more readable and free of grammatical errors. The sentences are better structured and organized as a whole paragraph.

Please feel free to add reasons that are not listed above (optional).

Supplementary Figure 1. User interface for collecting human feedback. The upper box shows an example summary produced by a studied LLM. The lower box displays the multiple-choice question about the rationale of the human evaluators' preference.



Supplementary Figure 2. Pearson Correlation Coefficients (r) among evaluation metrics. The natural language generation (NLG) metrics have a strongly positive correlation between each other (r >0.68). The PICO metrics have a moderate positive correlation with NLG metrics, (0.15<r<0.46). Recall that NLG metrics focus on lexical similarity while PICO metrics focus on coverage of key information (PICO elements in the summary).

Supplementary Table 1: Dataset used for medical evidence summarization task.

Dataset	n
Training	7,472
Validation	394
Test	295

Supplementary Table 2. Automatic evaluation scores of LLMs.

	BART	**GPT-3.5**	**PRIMERA**		**LongT5-base**		**LongT5-xl**		**Llama-2**	
Metrics	FT	ZS	ZS	FT	ZS	FT	ZS	FT	ZS	FT
ROUGE-L	17.74	23.15	18.90	20.48	14.67	23.66	14.72	**24.61**	17.21	19.91
METEOR	27.49	**28.83**	25.15	26.50	14.70	25.93	15.06	28.27	19.69	25.02
CHRF	**40.54**	39.74	39.25	37.84	22.24	36.38	22.99	38.81	30.24	36.42
PICO Precision	42.29	48.61	34.86	49.18	52.61	53.32	49.73	53.76	50.83	**55.28**
PICO Recall	59.63	**66.40**	56.88	49.77	31.77	54.36	31.25	60.21	45.58	48.97
PICO F1	49.49	56.41	43.22	49.47	39.61	53.83	38.38	**56.80**	48.07	51.93

Supplementary Table 3. Human evaluation of fine-tuned models

Model	AD/Dementia/ Neurology	Gastroenterology	Internal Medicine	Nephrology	Rheumatology	Surgery	Overall
PRIMERA	12/18	13/18	24/36	9/18	5/18	6/18	(69/126) 54.76%
LongT5	12/18	12/18	25/36	13/18	6/18	7/18	(75/124) 59.52%
Llama-2	15/18	10/18	24/36	8/18	9/18	8/18	(75/126) 58.73%

Supplementary Table 4. Simulated evaluation by GPT-4.

Model	Before Cutoff	After Cutoff
Llama-2		
ZS	50.00*	50.00*
FT	67.39	69.49
PRIMERA		
ZS	19.08	18.98
FT	55.56	54.58
LongT5		
ZS	35.02	31.53
FT	74.88	77.97

Supplementary Table 5. The number of summaries with better qualities.

Model	Desired Properties	# of summaries with better qualities			
		generated by the zero-shot Llama-2		generated by fine-tuned models	
		n	%	n	%
PRIMERA (ft)	Comprehensiveness	19	(15.08%)	47	(37.30%)
	Consistency	32	(25.40%)	45	(35.71%)
	Readability	38	(30.16%)	37	(29.37%)
	Specificity	33	(26.19%)	35	(27.78%)
LongT5-xl (ft)	Comprehensiveness	22	(17.46%)	38	(30.16%)
	Consistency	30	(23.81%)	49	(38.89%)
	Readability	32	(25.40%)	48	(38.10%)
	Specificity	26	(20.63%)	35	(27.78%)
Llama-2 (ft)	Comprehensiveness	23	(18.25%)	50	(39.68%)
	Consistency	34	(26.98%)	40	(31.75%)
	Readability	36	(28.57%)	52	(41.27%)
	Specificity	24	(19.05%)	35	(27.78%)

Supplementary Table 6. Summaries generated by zero-shot, fine-tuned LongT5, GPT-3.5-turbo,input article doi: 10.1002/14651858.CD012458.

Model	Summary
LongT5(zero shot)	Our secondary objective was to assess the effect of single-room isolation or cohorting, or both on neonatal mortality and perceived or documented adverse effects in newborn infants admitted to the NICU.
LongT5(fine-tuned)	We found no evidence to support or refute the use of single-room isolation or cohorting for preventing transmission of HAIs or colonization with HAI causing pathogens in newborn infants less than six months of age admitted to the neonatal intensive care unit (NICU).
GPT-3.5-turbo	This systematic review aimed to evaluate the impact of single-room isolation or cohorting, or both, on the transmission of healthcare-associated infections (HAIs) in newborn infants admitted to the neonatal intensive care unit (NICU). However, no published or ongoing trials were found to include in the review. Therefore, no conclusions could be drawn regarding the effectiveness of these isolation measures on preventing HAIs or colonization with HAI-causing pathogens in this population. Further research is needed in this area.
Human Expert (Reference)	The review found no evidence from randomized trials to either support or refute the use of patient isolation measures (single-room isolation or cohorting) in neonates with HAIs. Risks secondary to infection control measures need to be balanced against the benefits of decreasing horizontal transmission in the neonatal unit for optimal neonatal outcomes. There is an urgent need to research the effectiveness of patient isolation measures for preventing the transmission of HAIs in neonatal units. Well-designed trials randomizing clusters of units or hospitals to a type of patient isolation method intervention are warranted.