***Supplementary Table S5**: Modified QUADAS-AI*

Domain	**Subject selection**	**Index text (LLM)**	**Reference standard**	**Work-flow**
Signaling questions	Q1: Were the input data well described (source, size, and quality) Q2: Were the subject selection criteria clearly presented? Q3: Was data pre-processing performed?	Q4: Were the split and justification of training and test datasets provided? Q5: Was external verification performed?	Q6: Was the reference standard likely to correctly perform the intended task?	Q7: Was the LLM's performance evaluated in one of the following 1) real-time 2) near-real-time medical settings, 3) pre-existing clinical cases?
Description of risk of bias concerns	The quality assessment evaluates the risk of bias as low, high, or unclear. In case all signaling questions were answered yes, the risk of bias is low. In case any signaling question was answered no, the potential for bias is highlighted for further discussions by reviewers to evaluate the risk of bias. In case there is insufficient data reported by a reviewed study, the risk of bias is judged as unclear.			