Supplementary Methods: Response-Generation Prompt, Evaluation Rubric and CHART Checklist

Case-specific clinical content (patient information, consultation request notes and full model responses) is not included in this supplement to protect patient confidentiality.

Supplementary Figures

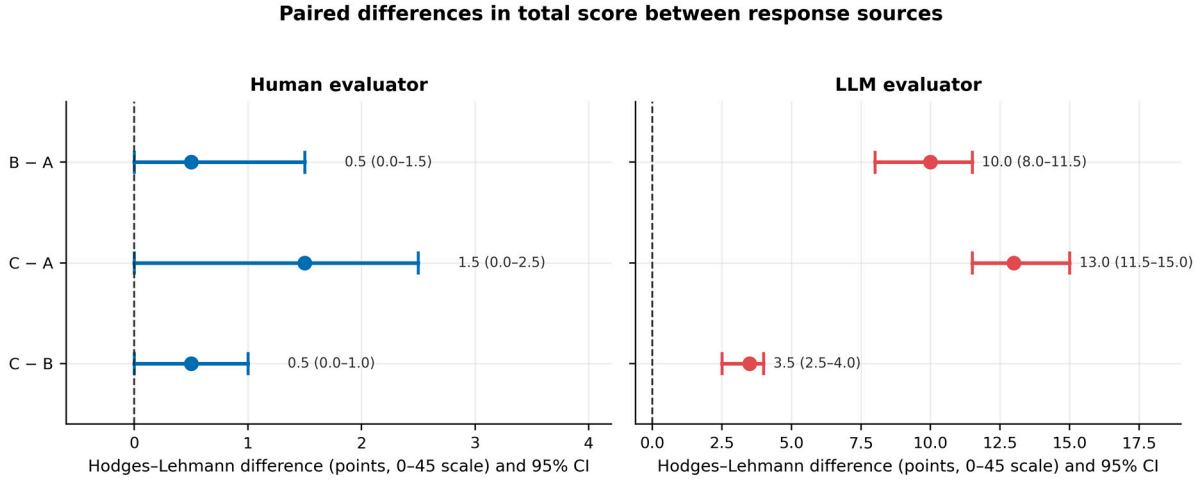


Figure S1. Paired differences in total score between response sources (Hodges–Lehmann estimates with 95% confidence intervals). Positive values favour the first-named source.

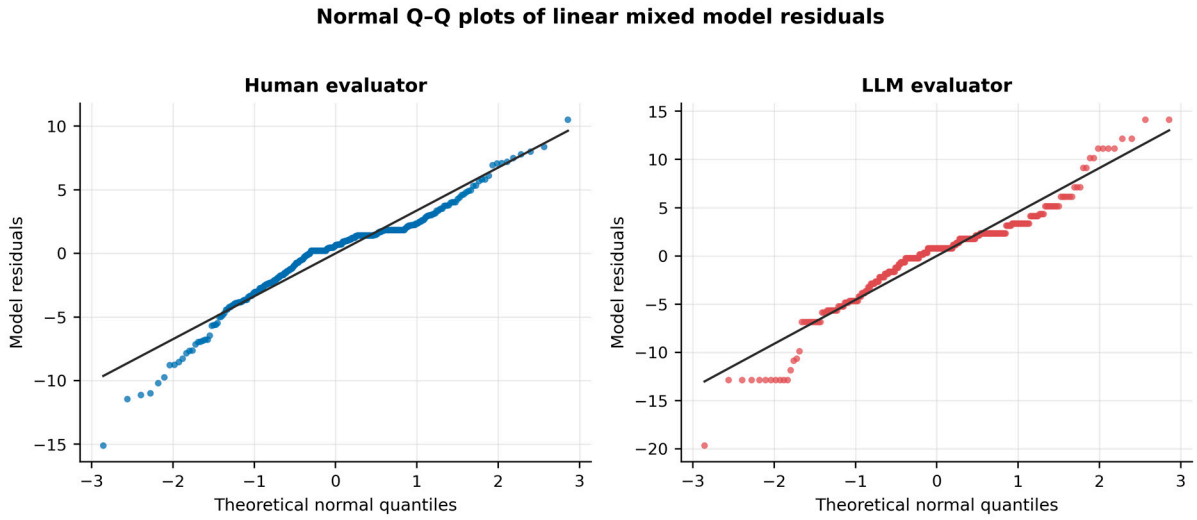


Figure S2. Normal Q–Q plots of linear mixed model residuals.

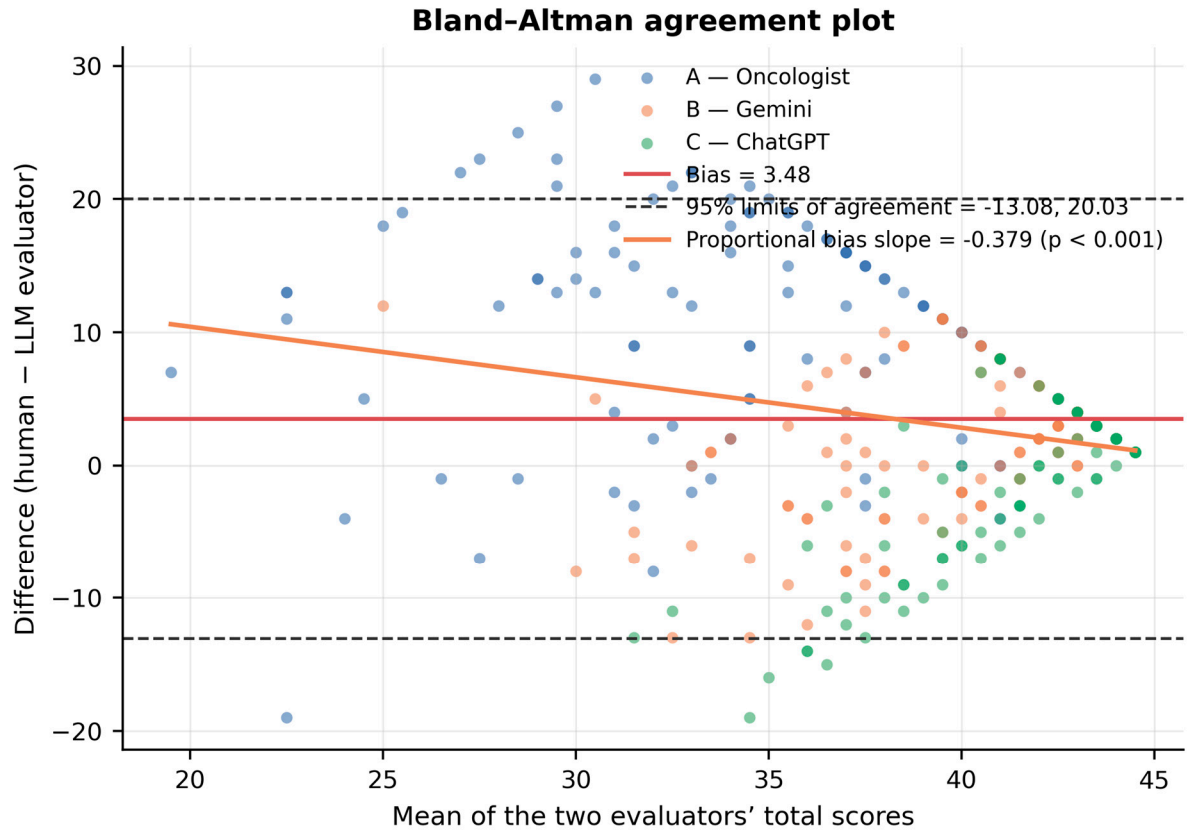


Figure S3. Bland–Altman plot of agreement between the two evaluators.

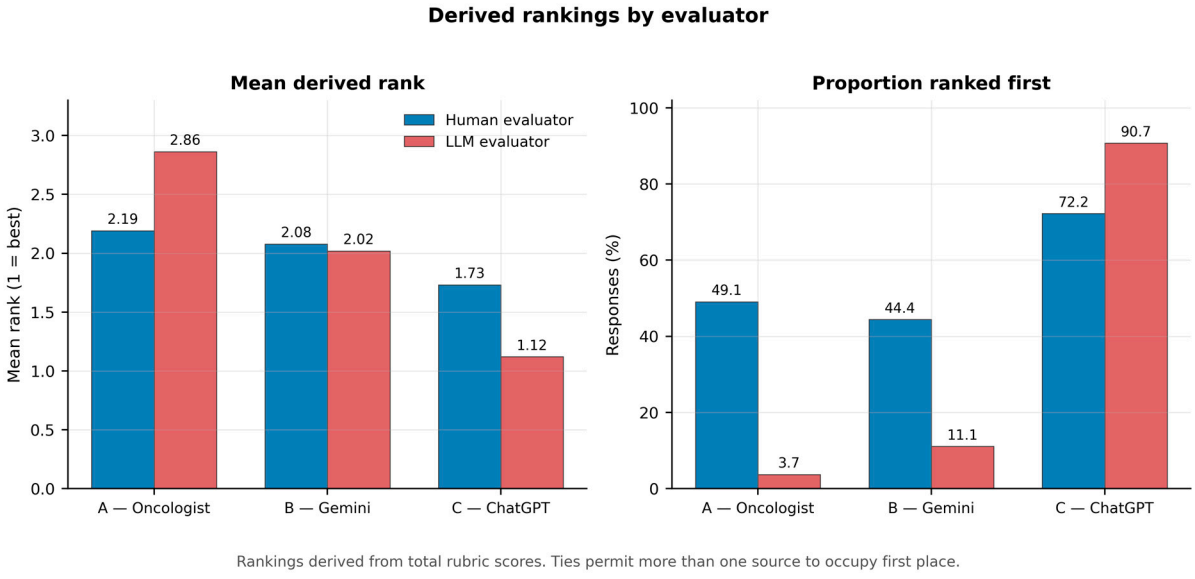


Figure S4. Mean derived rank and proportion ranked first, by evaluator.

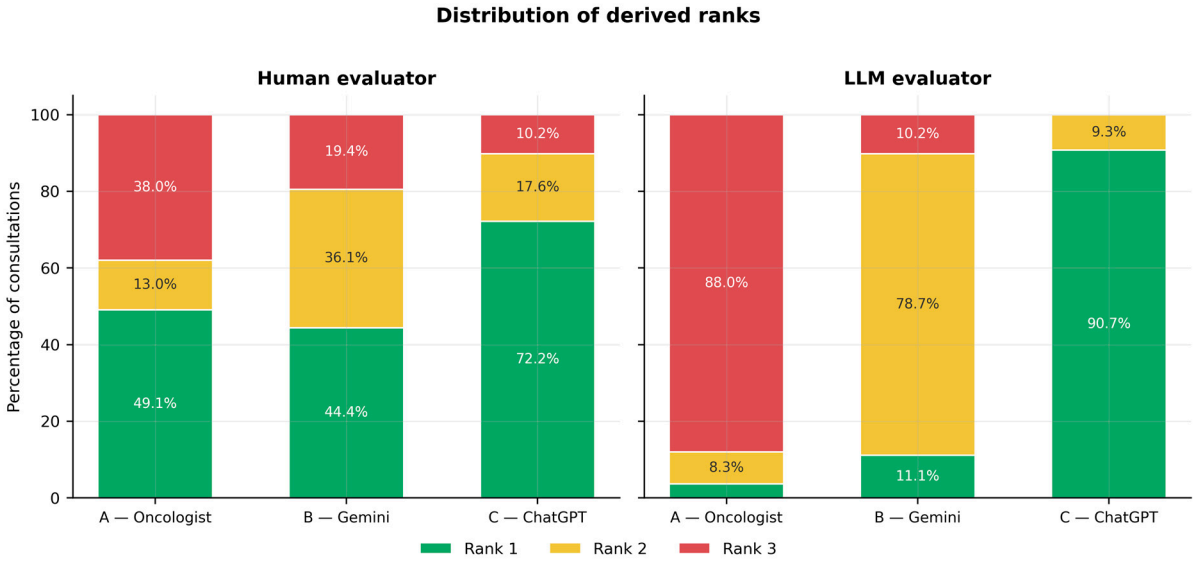
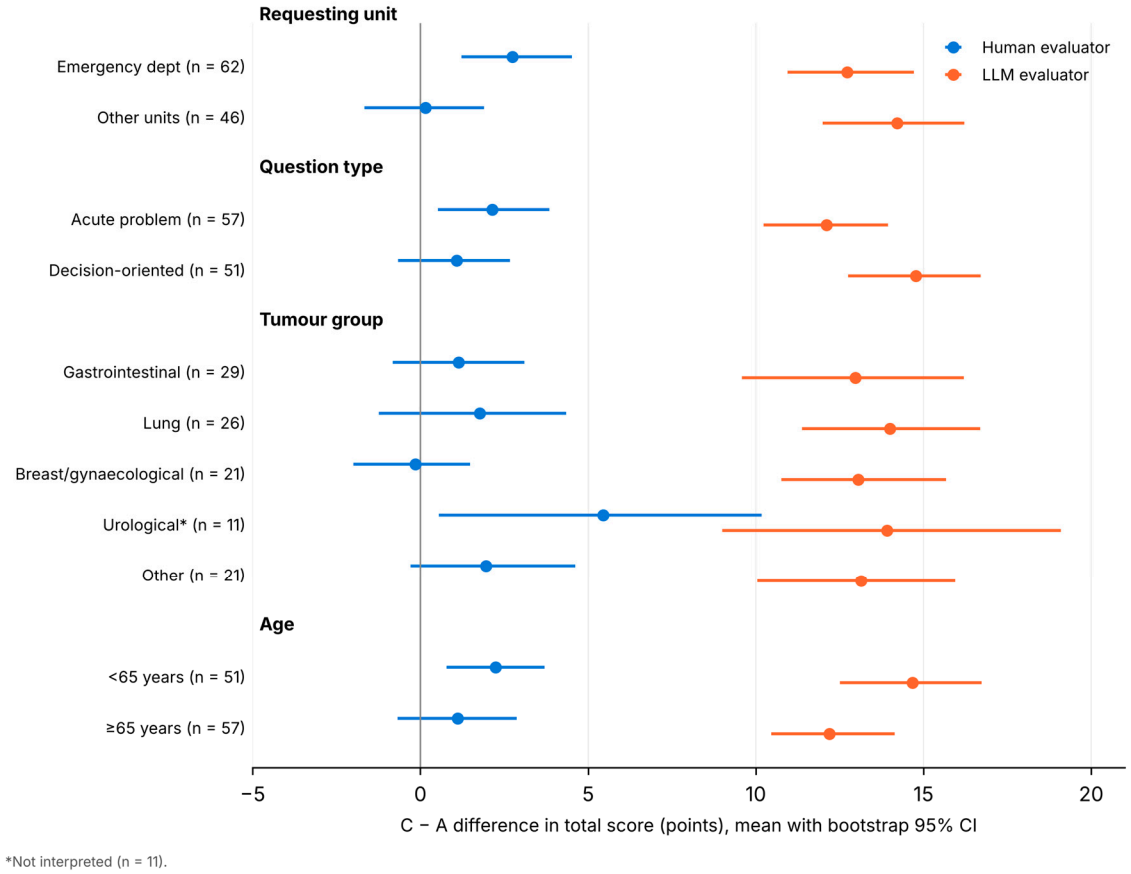


Figure S5. Distribution of derived ranks by response source and evaluator.



*Not interpreted (n = 11).

Figure S6. Exploratory subgroup analyses, by evaluator. Within-consultation difference in total score between source C and source A for each subgroup stratum, shown separately for the human and LLM evaluators (mean with bootstrap 95% confidence interval). The urological tumour stratum (n = 11) was not interpreted.