Benchmarking Large Language Models in Evidence-Based Medicine

Jin Li, Yiyan Deng, Qi Sun, Junjie Zhu, Yu Tian, Jingsong Li, and Tingting Zhu

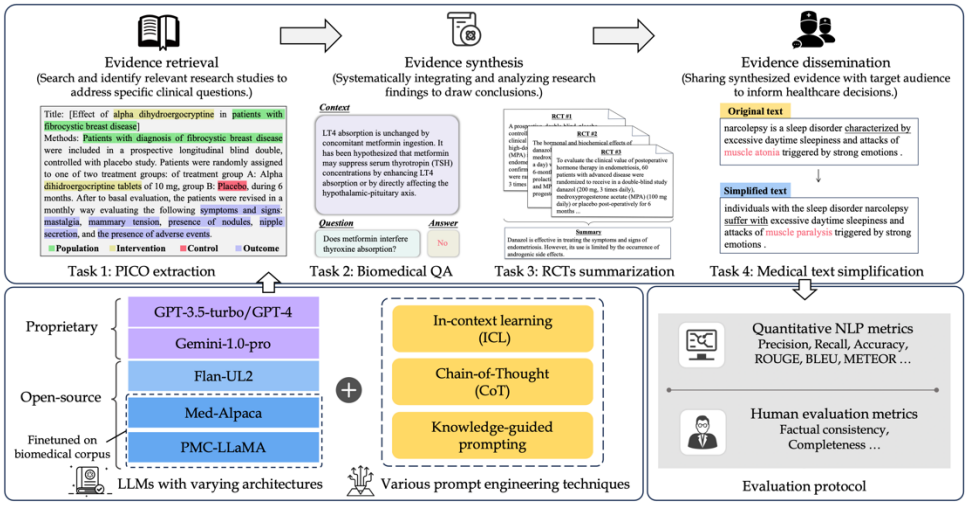

Fig. 1: The workflow of this study.

TABLE VII: An overview of error types.

Category	Source	Example
Factual contradiction	Reference summary	Post extubation atelectasis may be more frequent after **nasal intubation**, particularly in very low birth weight infants. One route of intubation does not seem to be preferable to the other...
	Output summary	**Flan-UL2-20b**: The authors compared the incidence of post-extubation atelectasis in patients intubated with oral versus nasal endotracheal tubes. The incidence of atelectasis was significantly higher in patients **intubated with oral tubes**.
Certainty Illusion	Reference summary	Metronidazole, given as a single dose, is likely to provide a parasitological cure for trichomoniasis, but it is **not known** whether this treatment will have any effect on pregnancy outcomes...
	Output summary	**gpt-3.5-turbo-1106**: In summary, the evidence from these studies suggests that treatment of asymptomatic trichomoniasis in pregnant women with metronidazole **does not** prevent preterm delivery, and the impact of Trichomonas infection on perinatal outcomes may be limited.
Evidence omission	Reference summary	Insufficient evidence exists from randomised trials to confirm the relative effects of **open versus closed reduction of intracapsular fractures**, or the effects of **intra-operative impaction or compression** of an intracapsular fracture treated by internal fixation.
	Output summary	**Flan-UL2-20b**: The results of this study suggest that **closed reduction** is as effective as **open reduction** in the treatment of displaced intracapsular fractures of the neck of the femur in young adults.
Semantic distortion	Original text	We observed **unfavourable generation shifts in diabetes** among men but not among women.
	Simplified text	**gpt-3.5-turbo-instruct**: We noticed that men were **more likely to develop diabetes** compared to women.
Fabricated attribute	Original text	Framingham offspring study participants who attended the seventh examination were at least 60 years old and free of dementia and stroke were included.
	Simplified text	**gpt-4-0613**: The Framingham Offspring Study included participants who were at least 60 years old. These participants did not have dementia or **a history of** stroke.
Misinterpretation	Original text	**Subordinate individuals** had lower heart rates than dominants, and consequently they suffered lower rates of body mass loss during periods of restricted pellet food supply.
	Simplified text	**Gemini-1.0-pro**: When **animals were lower in the social hierarchy**, their heart rates were slower. This helped them lose less weight when food was scarce.
Hallucination	Original text	Five patients with DM2 and stage II/III HF consumed dark chocolate and a beverage containing approximately 100 mg of EPI per day for 3 months.
	Simplified text	**PMC-LLaMA-13b**: The patients with DM2 and stage II/III HF who consumed dark chocolate and a beverage containing approximately 100 mg of epicatechin per day for three months **experienced a significant improvement in their cardiac function.**

TABLE II: Task-specific prompt used for querying models. Blue text denotes domain-specific prefixes. Yellow, blue and green highlights represent ICL demonstrations, knowledge-guided prompting, and CoT prompting, respectively. Input variables are denoted by red text enclosed in <>.

Task	Prompts
PICO extraction	**Task Description** You are a skilled medical expert. Your task is to generate an HTML version of an input text, marking up specific entities related to healthcare. The entities to be identified are: "Participant", "Intervention", "Control", and "Outcomes". Use HTML tags to highlight these entities. Each should have a class attribute indicating the type of the entity. **Markup Format** Use to denote a participant entity. Use to denote an intervention entity. Use to denote a control entity. Use to denote an outcome entity. Leave the text as it is if no such entities are found. **Entity Recognition Guide** [entity recognition guide for PICO extraction task] **Examples (N-shot)** [i=1,2,...N] Input i: **<example_input_text_i>** Annotated output i: **<example_output_i>** **Input Text** **<input_text>**
Biomedical QA	**Task Description** You are a skilled medical expert. Considering the information from a biomedical study provided in the reference text, is it correct to conclude that "**<question>**"? Please respond with "yes" or "no". Please first respond with "yes" or "no", followed by a brief explanation of your reasoning process, ensuring that your explanation aligns with the study's findings. **Examples (N-shot)** [i=1,2,...N] Question i: **<example_question_i>** Reference text i: **<example_reference_text_i>** Answer i: **<example_answer_i>** **Input Text** Question: **<question>** Reference text: **<reference>**
RCT summarization	**Task Description** You are a skilled medical expert. Consolidate the information from these randomized controlled trial abstracts into a comprehensive summary. **Summarization Guide** [summarizaiton guide for RCT summarization task] **Examples (N-shot)** [i=1,2,...N][j=1,2,... # References in Example i] Reference Text **<j>** in Example **<i>** Title: **<example_title_ij>** Abstract: **<example_abstract_ij>** Summarization **<example_summarization_i>** **Input Text** [j=1,2,... # References in Input] Reference Text **<j>** Title: **<title_j>** Abstract: **<abstract_j>**
Medical text simplification	**Task Description** You are a skilled medical expert. Simplify the given medical text, making complex medical information accessible and relevant to people without a medical background. **Summarization Guide** [summarizaiton guide for medical text simplification task] **Examples (N-shot)** [i=1,2,...N] Input i: **<example_input_text_i>** Simplified output i: **<example_output_i>** **Input Text** **<input_text>**