Supplementary Materials

MedFactEval and MedAgentBrief: A Framework and Workflow for Generating and Evaluating Factual Clinical Summaries

Model	**License**
gemini-2.5-pro-preview-05-06	Proprietary
gemini-2.0-flash-001	Proprietary
gpt-4.1-2025-01-01-preview	Proprietary
gpt-4o-2024-05-15	Proprietary
claude-3.7-sonnet-2025-02-19-v1.0	Proprietary
claude-3.5-sonnet-2024-10-22-v2.0	Proprietary
llama-4-maverick-17B-128E-instruct-FP8	Llama 4 license
llama-4-scout-17B-16E-instruct	Llama 4 license
deepseek-r1-original-release	MIT license
phi-3.5-mini-instruct	MIT license

Table S1: **Foundation Models Used for Generation and Evaluation.** This table lists the foundation models and their respective licenses used to implement the generation strategies (Single-Prompt and MedAgentBrief) and to serve as judges in the MedFactEval LLM Jury. Model versions are specified as provided by the API endpoints at the time of the study.

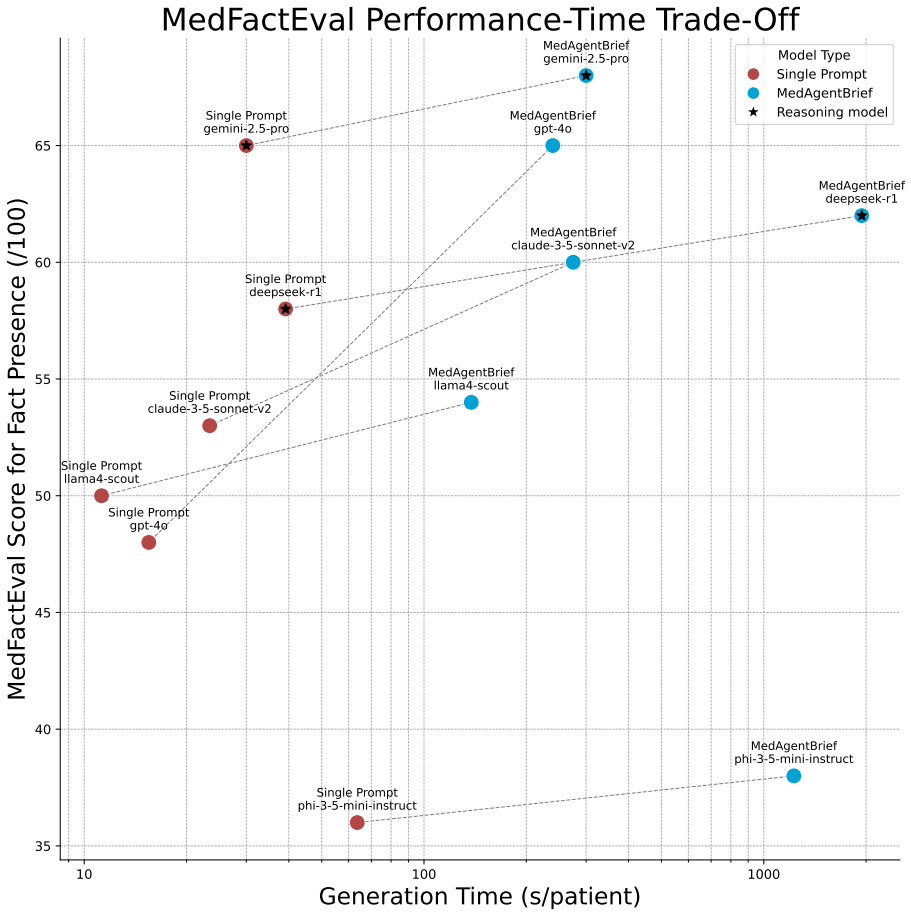

Figure S1: **Performance-Time Trade-off of AI Summary Generation, Measured by MedFactEval.** The y-axis represents the MedFactEval score for factual presence (higher is better), while the x-axis shows inference cost per patient on a logarithmic scale. For each foundation model, the MedAgentBrief workflow (cyan) consistently yields a higher factuality score than the baseline Single Prompt approach (brown), demonstrating its effectiveness at the expense of increased cost. Models designated as "Reasoning models" (starred) generally occupy the high-performance, high-cost quadrant.

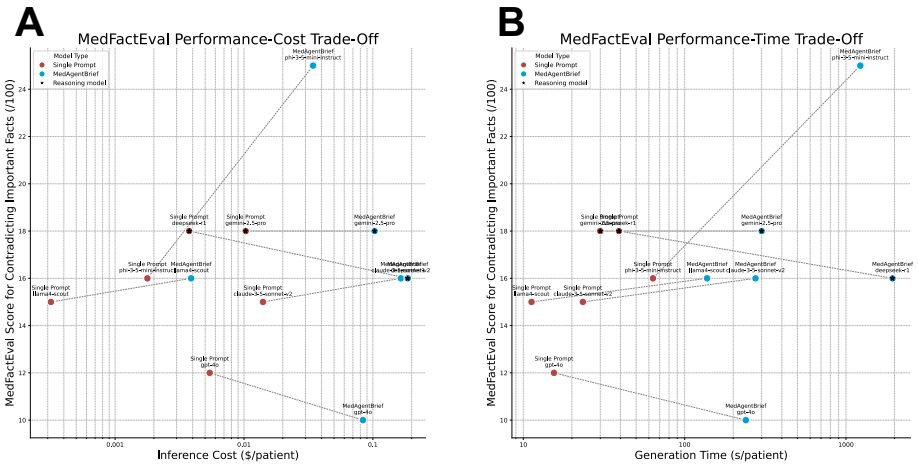

Figure S2: **Analysis of Contradiction Errors vs. Cost and Time, Measured by MedFactEval.** The y-axis represents the MedFactEval score for contradicting key facts (a lower score indicates fewer contradictions and is better). The x-axis shows inference cost (Panel A) and generation time (Panel B), both on a logarithmic scale. Unlike with fact presence, the impact of the MedAgentBrief workflow (cyan) on contradiction rates varies by model compared to the Single Prompt approach (brown). For some models (e.g., GPT-4o, Claude 3.5), MedAgentBrief reduces contradictions, while for others it has a neutral or slightly negative effect.

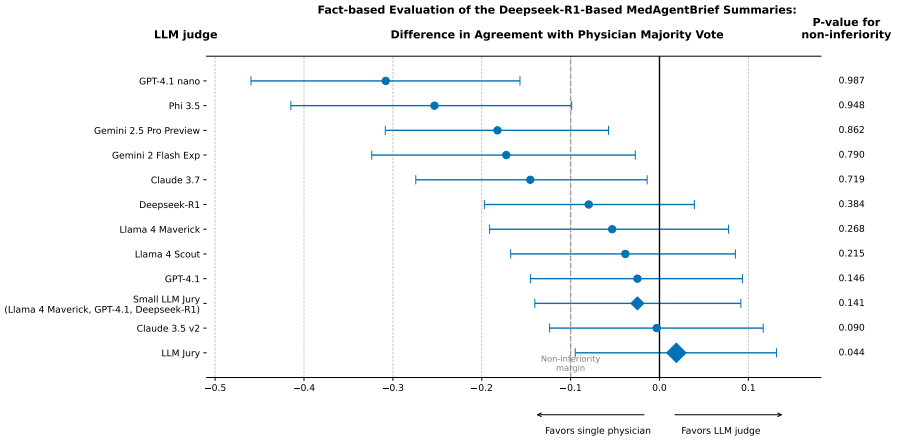

Figure S3: **Non-Inferiority of LLM Judges Compared to a Single Physician Expert.** The plot shows the difference in Cohen's kappa between each LLM judge and the average single physician, for summaries generated by the DeepSeek R1-based MedAgentBrief. Points to the right of the vertical line favor the LLM judge. The dashed line indicates the pre-specified non-inferiority margin. The full LLM Jury (bottom diamond) surpassed the single physician baseline and met the criteria for non-inferiority ($P < 0.05$). Error bars represent 90% confidence intervals, consistent with the one-sided nature of the non-inferiority test.