A Prompt Templates

All prompts used in this study are presented below.

A.1 Empathy Enhancement Prompts

Simple Editing Prompt

You will be provided with a patient's question and a physician's response. Your task is to revise the physician's response to make it more empathetic while preserving its original meaning and writing style. Patient Question: {PQ} Physician's response: {PR}

Fig. 1: Editing prompt for empathy enhancement. PQ refers to the patient question and PR refers to the physician's response.

Refine Editing Prompt

Key editing principles: 1. Preserve factual and clinical accuracy.
- Do not introduce or infer medical facts or conditions not explicitly stated by the physician.
- Do not make any clinical assumptions or diagnoses that are not present in the original response.
2. Respect the physician's intent.
- Do not add follow-up recommendations or next-step suggestions unless they already appear in the physician's original response.
- Do not add new advice, warnings, or treatment instructions.
3. Maintain emotional balance.
- Do not add false reassurance or overconfidence.
- Do not introduce unnecessary doubt, fear, or alarming language.
- Empathy should be expressed through tone, acknowledgment, and understanding—not through added medical content.
4. Preserve structure and style.
- Keep the response roughly the same length as the original.
- Maintain the physician's professional tone and sentence structure where possible.
- Revise only what's necessary to make the tone warmer, more understanding, or more supportive.

Fig. 2: The refined editing prompt with explicit behavioral constraints.

A.2 Evaluation Prompts

Empathy Evaluation Prompt

You are an expert in comparing the empathy level of two responses to a patient question. You will be given a patient question, and two responses, your task is to evaluate which answer is more empathetic.
Patient Question: {PQ}
Response 1: {R1}
Response 2:{R2}
Which response is more empathetic? Response 1 or 2?
- Response 1: More empathetic
- Response 2: More empathetic
- Both responses are equally empathetic

Fig. 3: 3-EMRank metric Prompt.

Factuality Evaluation Prompt

You are an expert on natural language entailment.
Your task is to deduce whether premise statements entail hypotheses.
Return only '1' if the hypothesis can be fully entailed by the premise.
Return only '0' if the hypothesis contains information that cannot be entailed by the premise.
Generate the answer in JSON format with the following keys:
'entailment_prediction': 1 or 0, whether the claim can be entailed.
Only return the JSON-formatted answer and nothing else.

Premise: {premise}
Hypothesis: {hypothesis}

Here is the JSON-formatted answer:

Fig. 4: Prompt for entailment evaluation

You are extracting medical facts from a physician's response to a patient question. Please breakdown the PHYSICIAN'S RESPONSE into independent MEDICAL facts as a string delimited by "//" to separate the facts.

DO NOT include as facts:
- Pure emotional support (e.g., "I understand", "I'm here to help")
- Non-medical expressions of care (e.g., "We care about you")
- General conversational elements (e.g., "Glad you're doing well")
- Repetitions or acknowledgments of patient's reported symptoms (e.g., "I understand you have pain")

Example 1:
Note: "Don't worry, we'll get through this together. The chest X-ray shows pneumonia in the right lung. Start antibiotics twice daily for 7 days. You're in good hands."
Atomic facts:
The chest X-ray shows pneumonia. // The pneumonia is in the right lung. // Start antibiotics. // Take antibiotics twice daily. // Continue antibiotics for 7 days.

Example 2:
Note: "There is a dense consolidation in the left lower lobe."
Atomic facts:
There is a consolidation. // The consolidation is dense. // The consolidation is on the left. // The consolidation is in a lobe. // The consolidation is in the lower portion of the left lobe.

Do not include any other text, or say "Here is the list..."
Note: {text}

Fig. 5: Prompt for medical fact extraction

B Data

Physician responses are concise (mean 65.8±50.4 words) while patient questions are longer and more variable (mean 82.2±56.2 words; Supplementary Fig. 6). Since 91.4% of questions require patient-specific EHR data, our approach of editing existing physician responses offers a more realistic alternative to end-to-end QA models.

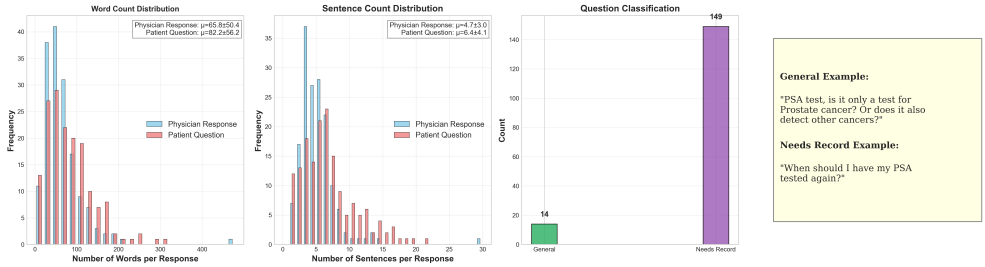

Fig. 6: Statistical characteristics of the patient-physician QA dataset (N=163). Physician responses are notably concise while patient questions show greater length variability. Question classification reveals that 91% require patient-specific EHR data rather than general medical knowledge.

C Metric Validation

C.1 3EM-Ranker Validation

We validated 3EM-Ranker through human evaluation with three prostate cancer patients comparing empathy between physician and AI responses. Alignment scores of 0.57 (Qwen-3) and 0.55 (LLaMA-3) substantially exceed the 0.23 reported for EM-Rank [1].

Table 1: Alignment scores between 3-EMRank and human annotators.

Model	Alignment Score
3-EMRank (LLaMA-3)	0.55
3-EMRank (Qwen-3)	0.57
Prior work (EM-Rank)	*0.23*

C.2 MedFactChecking Validation

Automated Fact Extraction Validation

Two machine learning scientists independently verified the accuracy of facts flagged as added (hallucinated, $[d \models c'] = 0$) or not preserved (missed, $[d' \models c] = 0$) by the algorithm. Validation achieved 83.6% precision for detecting hallucinated facts and 72.7% for not-preserved facts (see Supplementary Table 2).

Table 2: Precision of automated fact extraction validated by ML expert annotators.

Response Type	Total Facts	Flagged	Confirmed	Precision
Original Response	934	191 (Not Preserved)	139	72.7%
Edited Response	1230	262 (Added)	219	83.6%

Clinical Expert Evaluation

Two board-certified urologists from Mayo Clinic reviewed empathy-enhanced responses without access to the automated analysis, flagging 32 of 163 (19%) as containing potential fabrications or clinical inaccuracies. These errors fell into six patterns: follow-up recommendations (13 cases), clinical assumptions/speculation (7), clinical inaccuracies (4), unnecessary advice (4), unnecessary doubt/fear (2), and false assurance (2). To measure alignment, ML researchers mapped automatically extracted facts ($[d \models c'] = 0$ and $[d' \models c] = 0$) to these categories and computed coverage: the proportion of expert-identified fabrications where the algorithm detected at least one corresponding fact. MedFactChecking achieved 90.62% overall coverage (29/32), with perfect detection for clinical assumptions, unnecessary advice, and unnecessary doubt/fear, indicating that despite moderate precision in granular fact-level validation, the algorithm captures the majority of clinically impactful errors (see Supplementary Table 3).

Table 3: Expert-identified fabrication patterns and MedFactChecking coverage.

Pattern	Total (Expert)	Detected (Algorithm)	Missed	Coverage
Added follow-up recommendation	13	12	1	92.3%
Clinical assumption/speculation	7	7	0	100%
Clinical inaccuracy	4	3	1	75.0%
Adds unnecessary advice	4	4	0	100%
Adds unnecessary doubt/fear	2	2	0	100%
False assurance	2	1	1	50.0%
Overall	**32**	**29**	**3**	**90.62%**

D Additional Experiment Results

D.1 Factuality Evaluation Results

Table 4 quantifies fact transformation during empathy editing. All models add substantial content (28–53% increase in fact count), but differ dramatically in how much of this content is grounded in the original response. The results reveal two distinct failure modes. First, *information loss* (lower recall): non-Gemini models fail to preserve 20–22% of original medical facts, that might also include critical clinical information. Second, *hallucinated additions* (lower precision): these same models introduce 260–355 unsupported facts (21–25% of edited content), risking patient safety through fabricated medical claims. Gemini-2.5-Flash's superior performance stems from both minimal loss (8.5%) and controlled addition (9.5% hallucination rate).

Table 4: Fact flow analysis across empathy-editing models. Loss Rate measures information omission from original responses; Hallucination Rate measures unsupported additions in edited responses. Metrics are computed as: New Facts = Edited − Grounded; Loss Rate = (Original − Preserved) / Original; Hallucination Rate = New / Edited.

Model	**Original** $\lvert C\rvert$	**Preserved** $\sum \mathbb{I}[d' \models c]$	**Edited** $\lvert C'\rvert$	**Grounded** $\sum \mathbb{I}[d \models c']$	**New Facts**	**Loss Rate**	**Halluc. Rate**
Gemini-2.5-Flash	934	855	1,194	1,081	113	**8.5%**	**9.5%**
Llama-3.1-70B	934	743	1,230	970	260	20.4%	21.1%
Mistral-7B	934	732	1,373	1,053	320	21.6%	23.3%
Qwen3-14B	934	744	1,429	1,074	355	20.3%	24.8%

D.2 Relation of Empathy and Factuality

We evaluated three strategies using Gemini-2.5-Flash on general medical queries: (1) *direct AI-generated*, (2) *simple-prompt*, and (3) *refined-prompt*. Refined-prompt achieves near-perfect factual preservation (micro/macro F1: 0.96/0.93), simple-prompt shows modest degradation (F1: 0.92/0.88), while direct AI-generated fails catastrophically (F1: 0.39/0.32), preserving fewer than 40% of physician facts (Supplementary Fig. 7). Empathy evaluation using 3EM-Ranker reveals an inverse pattern (Supplementary Table 5): **Factuality**: physician > refined-prompt > simple-prompt > direct AI-generated; **Empathy**: direct AI-generated > simple-prompt > refined-prompt > physician. This demonstrates that editing physician responses achieves 93–99% factual preservation while improving empathy, whereas direct generation introduces unverifiable content unsuitable for clinical deployment.

References

[1] Man Luo, Christopher J Warren, Lu Cheng, Haidar M Abdul-Muhsin, and Imon Banerjee. Assessing empathy in large language models with real-world

Table 5: 3EM-RANKER comparison results: percentage of pairs where system A is judged more empathic than system B, and vice versa, under two LLM judges. Use Gemini to generate the patient responses.

	Qwen3 Judge		LLaMA3 Judge	
Comparison (A vs B)	**A > B (%)**	**B > A (%)**	**A > B (%)**	**B > A (%)**
Direct AI (A) vs Simple-Prompt Edited (B)	71.4	0.0	85.7	14.3
Simple-Prompt Edited (A) vs Refined-Prompt Edited (B)	71.5	21.4	64.2	35.7
Refined-Prompt Edited (A) vs Physician (B)	57.1	0.0	92.8	0.0

A and B denote the two systems in each comparison. Percentages may not sum to 100 due to ties or unclassified cases.

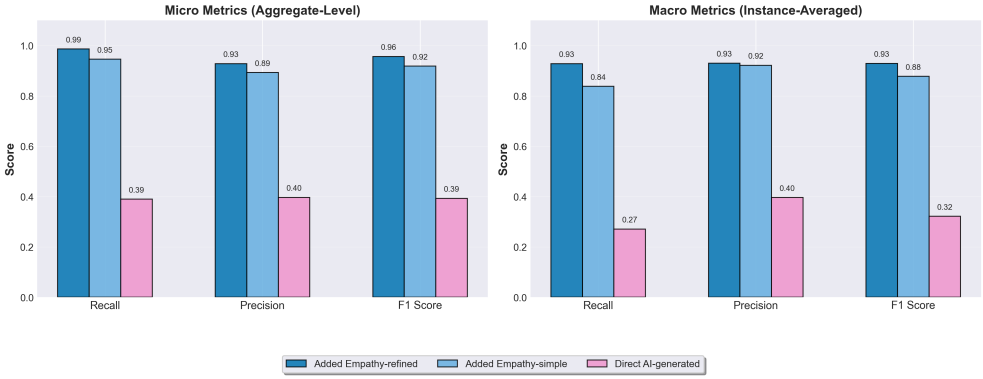

Fig. 7: MedFactChecking scores comparing three response generation strategies on general medical questions (n=14). Refined editing maintains high recall and precision (0.93–0.99), while direct AI generation fails to preserve physician facts (recall: 0.27–0.39), demonstrating the superiority of editing over generation for factual grounding.

physician-patient interactions. In *2024 IEEE International Conference on Big Data (BigData)*, pages 6510–6519. IEEE, 2024.