Prompt Design Shapes Clinical Detail In AI-Generated Low Back Pain Responses

Yale University
School of Medicine

Ahmed Basharat M.D[1], Omar Hamza M.D[2], • Paragi Rana M.D[1], Charles A. Odonkor M.D[1], Robert Chow M.D, MBA[1]

[1]Department of Emergency Medicine, Yale New Haven Hospital, Yale School of Medicine, New Haven, CT
[2]Department of Psychiatry, Cambridge Health Alliance, Harvard Medical School, Cambridge, MA



Introduction

Large language models (LLMs) are increasingly used for clinician support and patient education. In low back pain, the level of detail generated by an LLM may change depending on how the prompt is written.

Prior work suggests simplified prompts can produce shorter responses with less clinical detail. It remains unclear whether this reflects true information loss or prompt-driven suppression of knowledge.

This study evaluated whether clinically important information omitted under a fourth-grade prompt could be recovered by clinician-directed reframing.

Methods

Study design: Controlled comparative evaluation of LLM responses.

Questions: 15 standardized low back pain questions covering diagnosis, red flags, treatment, self-care, rehabilitation, and follow-up.

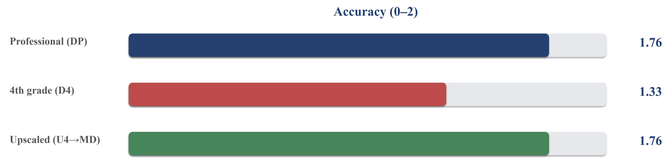
Model: ChatGPT GPT-5.0 (November 2025), single controlled session.

Three prompt conditions:

- DP: professional clinical prompt
- D4: fourth-grade reading-level prompt
- U4→MD: D4 response rewritten for a clinician

Three clinicians independently rated accuracy from 0–2. Clinical completeness was assessed by consensus using predefined domains: differential diagnosis, red flags, management, and escalation/follow-up. Word count and Flesch–Kincaid Grade Level (FKGL) were also measured.

Figures



DP vs D4: p = 0.00086 | DP vs U4→MD: p ≈ 1.00



KEY RESULT

Clinician-directed reframing restored accuracy to the professional-prompt level.

Linguistic Measures

Prompt	FKGL	Words	Interpretation
DP	11.27	49.7	Professional baseline
D4	7.13	35.5	Shorter / simpler
U4→MD	13.63	55.9	More complex / longer

D4 was significantly less complex than DP (FKGL p = 0.00459; word count p = 0.00049). U4→MD was more complex than DP (FKGL p = 0.00971).

Results

Accuracy and completeness moved together across prompt conditions:

- DP produced the highest mean accuracy (1.76/2).
- D4 reduced mean accuracy to 1.33/2.
- U4→MD returned accuracy to 1.76/2.
- Clinical completeness was 80.0% with DP, 6.7% with D4, and 73.3% after upscaling.

- Simplified prompts may suppress contraindications, red flags, diagnostic uncertainty, and next-step management.
- Prompt design can materially affect how much clinical reasoning is surfaced.
- LLMs may be better evaluated with audience-appropriate prompts rather than a single fixed prompting style.
- Clinician-directed reframing may recover omitted detail before clinical use.

Limitations:
- Single model/version and single controlled session.
- Small sample of 15 standardized questions.
- Low back pain only; generalizability to other conditions is unknown.
- Clinical completeness was subjective despite predefined criteria.
- Fleiss' κ = 0.17, indicating low inter-rater reliability.

Conclusion

- Reduced clinical detail after simplification appears to reflect constrained output rather than irreversible knowledge loss.
- Clinician-directed reframing restored accuracy to the professional-prompt level.
- LLM performance depends not only on model capability, but also on prompt design and intended audience.
- Structured prompting may be important when integrating AI into pain practice or patient-facing education.

References

Singhal K, et al. Nature. 2023;620:172–180.
Mesko B. J Med Internet Res. 2023;25:e50638.
Basharat A, et al. Interv Pain Med. 2025;4(3):100636.
Kreiner DS, et al. Spine J. 2020;20(7):998–1024.
JAMA/Thyroid and text-simplification literature cited in manuscript.