Human-Guided Iterative Prompt Engineering for Precision Feedback Message Authoring Using LLMs

MICHIGAN MEDICINE
UNIVERSITY OF MICHIGAN

Zhiyi Sun, Yidan Cao, Gan Shi, Allen Flynn, & Zach Landis-Lewis
University of Michigan Medical School

Introduction

- Healthcare professionals need to learn continuously as knowledge changes.
- Clinical performance feedback can aid learning but often lacks engagement.
- We developed a precision feedback system that enhances reporting systems with coaching and appreciation messages [1].

Precision Feedback

- Prioritizes **coaching** and **appreciation** messages.
- Uses estimates of the **motivational potential** of feedback messages.
- Supports performance **improvement** and **sustainment**.

Objective

Our goal: Explore the use of large language models (LLMs) to generate motivational messages (coaching, appreciation).

Why LLMs? Knowledge base development is complex and time-consuming. Generative AI may improve knowledge base development for message creation and metadata, but its effectiveness for these purposes is unclear [2].

Research Questions

- Can LLMs compose **acceptable** motivational messages?
- Can LLMs generate appropriate **metadata** for precision feedback messages?

Methods

LLMs Used: Mistral Large 2 and ChatGPT 4o

Procedure:

- 3 iterations of message generation.
- **Prompts**: Start simple, refine over iterations.
- **Message types**: 25 Coaching (improvement-focused) and 25 Appreciation (achievement-focused) for each LLM.
- **Evaluation**: We qualitatively assessed message correctness, consistency, and acceptability.

Figure 1. A precision feedback example for anesthetists from MPOG

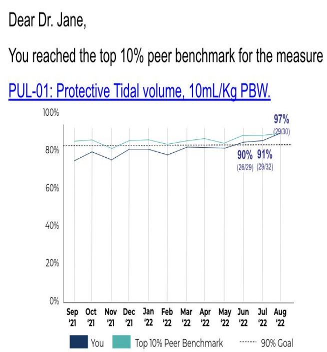

Figure 2. Examples of coaching and appreciation feedback messages

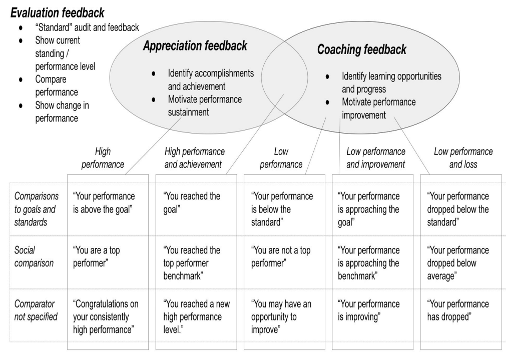

Figure 3. Human-guided iteration 1, 2, & 3

Iteration 1	Iteration 2	Iteration 3
• Initial prompt with minimal workflow instructions • Informal analysis of message acceptability	• Added definitions of key terms • Added metadata about message types and causal pathways	• Added detailed workflow instructions • Provided structured examples (accompanying with metadata)

Table 1. Messages composed categorized by motivating information

		Total messages composed	
Motivating information	**Motivating information subclass**	**ChatGPT**	**Mistral**
Comparisons	Better	0	1
	Worse	0	0
Trends	Improving	0	0
	Worsening	0	0
Status change	Achievement (Improving to a high level)	0	0
	Loss (Worsening to a low level)	3	3
Streak	Sustain high (remaining above a comparator with no apparent trend)	0	0
	Sustain low (remaining below a comparator with no apparent trend)	0	0
Anticipation of achievement	Approaching	15	4

Table 2. Messages composed categorized by comparison types

How many messages were specific to each comparator type?	ChatGPT	Mistral
Social comparison	10	6
Comparisons to goals and standards	8	2
Comparator not specified	0	0

Results

Iteration 1-3: Insights

- **Success Rates**:
  - ChatGPT: 36% success (coaching: 18, appreciation: 0).
  - Mistral: 16% success (coaching: 7, appreciation: 1).
- **Challenges**: Metadata inconsistencies, duplicates.
- **Key Takeaway**: ChatGPT created a wider variety of messages, but some were outside system scope.

Discussion

Observations:

- Refining prompts led to improved LLM performance.
- ChatGPT had better output variety but also irrelevant messages.
- Mistral had more duplicates, closely mirroring examples.

Key Insight: LLMs can generate useful messages, but further refinement is required.

Conclusion

Preliminary Insights:

- LLMs show potential but need prompt refinement and more advanced evaluation methods.
- Future directions:
  - Integrate **retrieval-augmented generation** techniques [3].
  - Develop stronger evaluation frameworks for message quality.
  - Share refined prompts as knowledge artifacts.

Reference

[1] Landis-Lewis, Z., Janda, A. M., Chung, H., Galante, P., Cao, Y., & Krumm, A. E. (2024). *Precision feedback: a conceptual model* (p. e10419).

[2] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, *29*(8), 1930-1940.

[3] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

https://github.com/Display-Lab/knowledge-base