

## **Zero-Shot Prompting is the Most Accurate and Scalable Strategy for Abstracting the Mayo Endoscopic Subscore from Colonoscopy Reports Using GPT-4**

Richard P. Yim<sup>1</sup>, MS; Vivek A. Rudrapatna, MD, PHD<sup>1,2</sup>

<sup>1</sup> Bakar Computational Health Sciences Institute, University of California – San Francisco, San Francisco, CA 94143

<sup>2</sup> Division of Gastroenterology and Hepatology, Department of Medicine, University of California – San Francisco, San Francisco, CA 94143

### **Financial/Grant Support**

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number K99LM014099, the National Center for Advancing Translational Sciences, National Institutes of Health, through UCSF-CTSI Grant Number UL1 TR001872, as well as the UCLA Clinical and Translational Science Institute through grant number UL1TR001881. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. This research project has benefitted from the Microsoft Accelerate Foundation Models Research (AFMR) grant program through which leading foundation models hosted by Microsoft Azure along with access to Azure credits were provided to conduct the research.

**Correspondence:**

Vivek A. Rudrapatna, MD, PhD

490 Illinois St, Ste 21E

San Francisco, CA 94143

Email: [vivek.rudrapatna@ucsf.edu](mailto:vivek.rudrapatna@ucsf.edu)

Fax: 415-502-4444

**Conflicts of Interest:**

VAR receives research support from Alnylam, Takeda, Merck, Genentech, Blueprint Medicines, Stryker, Mitsubishi Tanabe, and Janssen. He also is a shareholder of ZebraMD. RPY has nothing to disclose.

**Writing Assistance:**

None.

**Author Contributions:**

*Yim*: Study concept and design; data extraction; analysis and interpretation of data; drafting of manuscript; critical revision of the manuscript for important intellectual content

*Rudrapatna*: Study concept and design; data extraction; analysis and interpretation of data; drafting of manuscript; study supervision; critical revision of the manuscript for important intellectual content

### **Data Acknowledgement:**

The authors thank UCSF Academic Research Services for technical support related to enabling software in a secure, PHI compliant environment; UCSF AI Tiger Team for facilitating and managing access to Versa API (UCSF secure access to Microsoft Azure, OpenAI Large language Models); and the Chancellor's Task Force for Generative AI.

### **Word Count:**

1024 words.

### **Structured Abstract:**

**Introduction:** Large-language models can help extract information from clinical notes, making them potentially useful for research in ulcerative colitis. However, it remains unclear if these models will scale well in practice.

**Methods:** We analyzed the performance and cost of programmatically using GPT-4 to abstract Mayo endoscopic subscores (MES) from 499 colonoscopy reports using different prompting strategies.

**Results:** Zero-shot prompting, where GPT-4 is instructed without examples, was most accurate (83.55%) *and* cost-effective (\$0.097/note).

**Discussion:** Using GPT-4 to automatically curate the MES and other variables is a practical strategy for quantifying UC activity and measuring improvements to clinical care.

## **Introduction:**

The Mayo endoscopic subscore (MES) is a core measure of ulcerative colitis (UC) activity (1), but is not always explicitly documented in colonoscopy reports. Thus, clinical studies that use the MES frequently require manual review of these reports to abstract these scores from free-text descriptions. Large language models like GPT-4 have shown promise in their ability to extract information from clinical notes (2). Prior studies have used the more user friendly, chatbot interface to interact with these models. However, these models can also be used in a programmatic fashion, raising the possibility of being natively deployed within electronic health record (EHR) systems to dynamically maintain disease registries, optimize study recruitment, and support quality improvement.

As a next step, we studied the scalability and cost-effectiveness of using GPT-4 to automate the extraction of the modified MES. We hypothesized that more sophisticated, *n-shot* and iterative-style prompts would yield more accurate results despite higher costs associated with this strategy.

## **Methods:**

We utilized an existing set of 499 annotated colonoscopy reports sourced from two hospitals in California (3). 217 were from San Francisco General Hospital (SFGH), a safety net hospital, and 282 from the University of California, San Francisco Health, a tertiary care hospital. These reports were annotated based on 1) their suitability for MES scoring (e.g. clear diagnosis of UC, surgically unaltered anatomy), and, 2) the modified MES (4) if appropriate.

We developed two generic conversation templates for zero-shot and *n-shot* prompting to programmatically interact with GPT-4-turbo via LangChain (5), a framework that enables context-rich prompts, and UCSF Versa, a PHI-compliant programmatic interface with GPT-4. (See **Table 1**,

**Supplemental Digital Content 1**, for precise prompt templates and protocol texts.) We refer to  $n$ -shot prompting as providing  $n$  colonoscopy reports per Mayo score plus a non-Mayo scorable report in addition to the scoring protocol; zero shot refers to providing only the scoring protocol. For these templates we also studied the performance of GPT-4 with prompts that asked it to not only include the MES, but an *explanation* as well. We also provided GPT-4 with a *parsed* variation of colonoscopy report for UCSF and SFGH centers where only the main relevant text of the colonoscopy procedure report was provided as opposed to the colonoscopy report text in its entirety, which includes extraneous text strings. See **Method Details, Supplemental Digital Content 1**, for additional explanation.

### **Results:**

Zero-shot prompts on trimmed notes produced the best performing results consistently on both UCSF (81.45-83.55% weighted average accuracy) and SFGH (73.15-78.11% weighted average accuracy) reports (**Table 1**). We found that that  $n$ -shot prompting actually decreased classification performance across multiple metrics, rejecting our hypothesis that providing examples would improve GPT-4's performance. Further, the cost of  $n$ -shot prompting is prohibitively more expensive on average (e.g., more than *8 times* the cost for UCSF reports and *11 times* for SFGH reports between  $n$ -shot and zero-shot prompting *per note*). For Mayo scorable accuracy, whether an MES can be assigned to the colonoscopy report, we find that GPT-4 performs very well for zero-shot prompt templates (accuracy 90.28-93.66%), where  $n$ -shot prompting reduced its accuracy (best score at 84.28%). We also studied results for splitting Mayo scorable reports and MES separately ("zero-shot, two-task prompting") but performance gains were negligible (**Table 1**).

**Table 1.** Performance results, classification error, and cost are shown across each prompt template variation and data variation. Bolded text means higher is better; italicized text means lower is better. Best numeric values along each measurement have been underlined. Table has been partitioned according to template variation.

Prompt Template Variation	Zero-shot	Zero-shot	Zero-shot	Zero-shot	Zero-shot	Zero-shot	Zero-shot	Zero-shot	Zero-shot	Zero-shot, two-task	Zero-shot, two-task	One-shot	One-shot	Four-shot	Four-shot
Center	UCSF	UCSF	SFGH	SFGH	UCSF	UCSF	SFGH	SFGH	UCSF	SFGH	SFGH	UCSF	SFGH	UCSF	SFGH
With Parsing	Trimmed	Trimmed	Trimmed	Trimmed	Original	Original	Original	Original	Trimmed	Trimmed	Trimmed	Trimmed	Trimmed	Trimmed	Trimmed
With Explanation	Yes	No	Yes	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Accuracy*	<b><u>83.55%</u></b>	<b>82.16%</b>	<b>77.17%</b>	<b>78.11%</b>	<b>83.11%</b>	<b>81.45%</b>	<b>73.15%</b>	<b>75.86%</b>	<b><u>83.95%</u></b>	<b>78.59%</b>	<b><u>79.23%</u></b>	<b>73.04%</b>	<b><u>75.93%</u></b>	<b>73.82%</b>	
Precision*	<b><u>87.33%</u></b>	<b>86.94%</b>	<b>81.83%</b>	<b>82.72%</b>	<b>87.01%</b>	<b>85.37%</b>	<b>79.29%</b>	<b>81.37%</b>	<b><u>87.17%</u></b>	<b>82.57%</b>	<b><u>82.78%</u></b>	<b>80.51%</b>	<b><u>80.46%</u></b>	<b>77.88%</b>	
Recall*	<b><u>85.11%</u></b>	<b>84.75%</b>	<b>79.26%</b>	<b>80.18%</b>	<b>83.69%</b>	<b>82.27%</b>	<b>75.58%</b>	<b>78.34%</b>	<b><u>85.11%</u></b>	<b>81.11%</b>	<b><u>75.09%</u></b>	<b>74.06%</b>	<b><u>70.61%</u></b>	<b>74.62%</b>	
F1 score*	<b><u>85.62%</u></b>	<b>85.37%</b>	<b>79.51%</b>	<b>80.67%</b>	<b>84.51%</b>	<b>83.12%</b>	<b>76.42%</b>	<b>78.81%</b>	<b><u>85.42%</u></b>	<b>81.16%</b>	<b><u>76.42%</u></b>	<b>75.88%</b>	<b><u>71.95%</u></b>	<b>75.64%</b>	
Scorable Accuracy	<b>92.95%</b>	<b><u>93.66%</u></b>	<b>92.24%</b>	<b>92.57%</b>	<b>91.81%</b>	<b>90.28%</b>	<b>92.47%</b>	<b>91.81%</b>	<b><u>92.34%</u></b>	<b>92.33%</b>	<b><u>82.35%</u></b>	<b><u>84.28%</u></b>	<b><u>78.76%</u></b>	<b>83.33%</b>	
Severity Agreement †	<b>90.73%</b>	<b><u>91.39%</u></b>	<b>82.12%</b>	<b>82.78%</b>	<b>89.40%</b>	<b>90.73%</b>	<b>78.81%</b>	<b>82.12%</b>	<b><u>90.73%</u></b>	<b>82.78%</b>	<b><u>92.52%</u></b>	<b>80.95%</b>	<b><u>93.33%</u></b>	<b>82.22%</b>	
Under Classification(1)	<b><u>7.28%</u></b>	<b>8.61%</b>	<b>10.60%</b>	<b>11.92%</b>	<b>8.61%</b>	<b>9.27%</b>	<b>14.57%</b>	<b>12.58%</b>	<b><u>3.97%</u></b>	<b>7.95%</b>	<b><u>9.52%</u></b>	<b>15.66%</b>	<b><u>10.37%</u></b>	<b>12.59%</b>	
Under Classification(2)	<b><u>0.66%</u></b>	<b><u>0.66%</u></b>	<b>1.32%</b>	<b>1.32%</b>	<b><u>0.66%</u></b>	<b><u>0.66%</u></b>	<b>1.32%</b>	<b>1.32%</b>	<b><u>0.66%</u></b>	<b><u>0.66%</u></b>	<b><u>0.68%</u></b>	<b>1.36%</b>	<b><u>0.74%</u></b>	<b><u>0.74%</u></b>	
Average cost per note	<b>\$0.097</b>	<b><u>\$0.096</u></b>	<b>\$0.129</b>	<b>\$0.128</b>	<b>\$0.108</b>	<b>\$0.107</b>	<b>\$0.152</b>	<b>\$0.151</b>	<b><u>0.148</u></b>	<b>\$0.267</b>	<b><u>\$0.499</u></b>	<b><u>\$0.280</u></b>	<b><u>\$0.803</u></b>	<b>\$1.485</b>	

\*Computed using weighted-average of metric for each class.

† Severe UC as measured by having an MES of 2 or 3.

With respect to prompt variations such as parsing colonoscopy reports and soliciting explanations for MES values, across all strata we find that the greatest difference in performance is 4.02% for trimming and -3.31% for requiring an explanation (**Table 2**). In particular, parsing the text generally increases performance across all measures and decreases under classification as well. Interestingly, we find that across standard statistical learning metrics, the UCSF data shows an increase in performance when requiring explanation of the MES and a decrease in performance for SFGH data although the magnitude of these differences is minimal (worst difference in magnitude amongst accuracy, precision, recall and F1-score is 2.76%).

**Table 2.** Differences in performance stratified on parsing colonoscopy report text (*Parsed Report - Original Report*) and requiring explanation (*Prompt with Explanation - Prompt without explanation*).

	<i>(Parsed Report — Original Report)</i>					<i>(Prompt with Explanation — Prompt without Explanation)</i>			
<i>Center</i>	UCSF	UCSF	SFGH	SFGH	<i>Center</i>	UCSF	UCSF	SFGH	SFGH
<i>With Explanation</i>	Yes	No	Yes	No	<i>With Parsing</i>	Trimmed	Original	Trimmed	Original
Accuracy*	0.44%	0.71%	<b>4.02%</b>	2.25%	Accuracy*	1.39%	1.66%	-0.94%	<b>-2.71%</b>
Precision*	0.32%	1.57%	<b>2.54%</b>	1.35%	Precision*	0.39%	1.64%	-0.89%	<b>-2.08%</b>
Recall*	1.42%	2.48%	<b>3.68%</b>	1.84%	Recall*	0.36%	1.42%	-0.92%	<b>-2.76%</b>
F1 score*	1.11%	2.25%	<b>3.09%</b>	1.86%	F1 score*	0.25%	1.39%	-1.16%	<b>-2.39%</b>
Scorable Accuracy	1.14%	<b>3.38%</b>	-0.23%	0.76%	Scorable Accuracy	-0.71%	<b>1.53%</b>	-0.33%	0.66%
Severity Agreement†	1.33%	0.66%	<b>3.31%</b>	0.66%	Severity Agreement†	-0.66%	-1.33%	-0.66%	<b>-3.31%</b>
Under Classification(1)	-1.33%	-0.66%	<b>-3.97%</b>	-0.66%	Under Classification(1)	-1.33%	-0.66%	-1.32%	<b>1.99%</b>
Under Classification(2)	0.00%	0.00%	0.00%	0.00%	Under Classification(2)	0.00%	0.00%	0.00%	0.00%

\*Computed using weighted-average of metric for each class.

† Severe UC as measured by having an MES of 2 or 3.

### **Discussion:**

This is among the first few studies to use an LLM in a programmatic fashion to extract study variables from clinical notes. We found that the most accurate *and* scalable prompting strategy is conveniently the most simple when it comes to producing MES scores from colonoscopy reports. Zero-shot prompts are not only easy to implement, but cost effective. GPT-4 proves itself to be reasonably effective at being able to simultaneously determine whether a colonoscopy procedure report is Mayo scorable, and providing an MES when it is. Further, we find that *n-shot* prompting is unreliable both in performance and cost.

Beyond template variations of zero-shot,  $n$ -shot, and zero-shot, two-task prompting, our study explores prompt parameter interactions in GPT-4 performance that are currently absent in the literature (e.g., explanation requirement and text parsing). Further, our study explores the possibility of the generalizability of LLM information extraction across different centers. The primary limitation of our study then is a sophisticated endpoint. For instance, although the colonoscopy report distribution for IBD patients is representative across UCSF and SFGH centers, we have limited class representation for more severe MES graded UC (Mayo scores 2 and 3 in particular). Other studies in the literature explore continuously valued endpoints as well multidimensional endpoints extracted from clinical text (6,7).

However, these studies focus primarily on the performance of GPT-4 and LLMs on clinical notes. There has been little consideration and commentary on prompt engineering *and* consequently the costs of deployment—in other words, the practicality of deploying GPT-4 for other studies. Deploying four-shot prompts on 282 parsed UCSF colonoscopy reports, requiring GPT-4 to produce an explanation, comes out to an average total cost of \$418.77. There are thousands of colonoscopy procedure reports for IBD patients at our medical centers, but billions of notes across all diseases and patients in the US (8). Generative AI is and will be very expensive to deploy across all clinical areas and target variables. While LLMs like GPT-4 will enable retrospective information extraction, and consequently new observational studies using EHR data, we strongly advise clinical researchers to be mindful of various prompting strategies and their costs.



## **References:**

1. Lewis JD, et al. Use of the noninvasive components of the mayo score to assess clinical response in Ulcerative Colitis: *Inflammatory Bowel Diseases*. 2008;14(12):1660-1666.
2. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. *Radiology*. 2023;308(3):e231362.
3. Silverman AL, Bhasuran B, Mosenia A, et al. Accurate, Robust, and Scalable Abstraction of Mayo Endoscopic Subscores from Colonoscopy Reports. *Gastroenterology*. 2022;162(7):S-617-S-618.
4. Richards K, et al. *Ulcerative Colitis: Developing Drugs for Treatment Guidance for Industry*. USHHS, CEDR, CEBR. Federal Registrar; April 29, 2022.
5. Harrison C. *LangChain AI*. LangChain (Version 1.2.0) [Computer software]. <https://github.com/langchain-ai/langchain>. Released October 17, 2022.
6. Ge J, et al. *A Comparison of Large Language Model versus Manual Chart Review for Extraction of Data Elements from the Electronic Health Record*. *Gastroenterology*; 2023.
7. Williams CYK, et al. *Assessing Clinical Acuity in the Emergency Department Using the GPT-3.5 Artificial Intelligence Model*. *Health Informatics*; 2023.
8. Dinov ID. *Volume and value of big healthcare data*. *J Med Stat Inform*. 2016;4(1):3.