Journal of the American Medical Informatics Association, 2025, **32(7)**, 1164–1173
https://doi.org/10.1093/jamia/ocaf084
Advance access publication 3 June 2025
Research and Applications

Research and Applications

Dynamic few-shot prompting for clinical note section classification using lightweight, open-source large language models

Kurt Miller, MS[1,2], Steven Bedrick, PhD[3], Qiuhao Lu, PhD[4], Andrew Wen, MS[4], William Hersh, PhD[3], Kirk Roberts, PhD[4], Hongfang Liu, PhD[4,*]

[1]Bioinformatics and Computational Biology Program, University of Minnesota, Rochester, MN, United States, [2]Center for Digital Health, Mayo Clinic, Rochester, MN, United States, [3]Department of Medical Informatics & Clinical Epidemiology, Oregon Health and Science University, Portland, OR, United States, [4]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, United States

*Corresponding author: Hongfang Liu, PhD, McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, University of Texas Health, Houston, TX, United States (hongfang.liu@uth.tmc.edu)

Abstract

Objective: Unlocking clinical information embedded in clinical notes has been hindered to a significant degree by domain-specific and context-sensitive language. Identification of note sections and structural document elements has been shown to improve information extraction and dependent downstream clinical natural language processing (NLP) tasks and applications. This study investigates the viability of a dynamic example selection prompting method to section classification using lightweight, open-source large language models (LLMs) as a practical solution for real-world healthcare clinical NLP systems.

Materials and Methods: We develop a dynamic few-shot prompting approach to classifying sections where section samples are first embedded using a transformer-based model and deposited in a vector store. During inference, the embedded samples with the most similar contextual embeddings to a given input section text are retrieved from the vector store and inserted into the LLM prompt. We evaluate this technique on two datasets comprising two section schemas, including varying levels of context. We compare the performance to baseline zero-shot and randomly selected few-shot scenarios.

Results: The dynamic few-shot prompting experiments yielded the highest F1 scores in each of the classification tasks and datasets for all seven of the LLMs included in the evaluation, averaging a macro F1 increase of 39.3% and 21.1% in our primary section classification task over the zero-shot and static few-shot baselines, respectively.

Discussion and Conclusion: Our results showcase substantial performance improvements imparted by dynamically selecting examples for few-shot LLM prompting, and further improvement by including section context, demonstrating compelling potential for clinical applications.

Key words: section classification; clinical natural language processing; large language model; electronic health records.

Introduction

The copious patient medical data in unstructured text form compiled into Electronic Health Records (EHRs) is recognized as a predominantly untapped bounty of clinical information riddled with increasing complexity, quality, heterogeneity, length, and redundancy issues that greatly inhibit secondary usage.[1–8] The significant degree of domain-specific, context-sensitive language makes extracting structured clinical information from these data sources especially challenging.

Identification of note section boundaries and classes, as well as other document structural elements, has been shown to help disambiguate this confounding language, thereby improving information extraction and other clinical natural language processing (NLP) tasks and applications such as named entity recognition,[9] cohort retrieval,[10] event extraction,[11] medication extraction,[12] and abbreviation resolution.[13] The added section or structural information improves performance in these applications by providing context to concept mentions and relationships, allowing the models to filter or differentially weigh these concepts by their section constituency.

The conventional section schema widely taught to medical students and used by physicians to document a patient encounter, represented by the SOAP ("Subjective," "Objective," "Assessment," and "Plan") acronym, generally captures a patient's reported symptoms, a physical examination with lab measurements, the consulting physician's evaluation and a plan for treatment as well as any further assessment needed. In many institutions, alternative schemas such as HL7 Clinical Document Architecture (CDA)[14] are adopted from standardized ontologies. Compared to the SOAP schema, these alternative nomenclatures commonly contain more granular subsections such as an explicit review of each biological organ system, called a "Review of Systems" (for example). Although most providers conform to

Received: December 29, 2024; **Editorial Decision:** May 9, 2025; **Accepted:** May 13, 2025
© The Author(s) 2025. Published by Oxford University Press on behalf of the American Medical Informatics Association. All rights reserved.
For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Downloaded from https://academic.oup.com/jamia/article/32/7/1164/8155977 by Mayo Clinic Library user on 30 June 2025

similar clinical note syntax and composition, the overall structure of information across note types varies vastly between institutions, EHRs, and even the same providers over time,[7,15–17] so ascertaining section information from clinical notes is far from a trivial task.

Consequently, most previous attempts at automatic section classification involved highly engineered rule- or heuristic-based approaches tailored finely to a specific institution's note-taking practices. A 2019 systematic review on clinical note section identification found that existing section-annotated datasets used custom term dictionaries along with a standard terminology such as UMLS Metathesaurus, though a growing research movement was beginning to explore more traditional machine learning, transformer-based, and hybrid approaches.[18] More recently, transformer-based architectures and their high-parameter, large language model (LLM) agglomerations have accelerated advancements in state-of-the-art performance.[19] A recent application of a fine-tuned BioClinicalBERT model has been able to achieve a 0.973 F1 score on the in-domain section classification task for the MIMIC-III Progress Note SOAP-labelled dataset.[20]

However, superior to rule-based and statistical machine learning approaches in particular benchmark performance, supervised transformer and LLM fine-tuning methods still require large, domain-specific annotated corpora for model training and evaluation.[21] Even additional (also referred to as "continual") pretraining, while demonstrating noteworthy performance through supplementary domain- or task-specific self-supervision on clinical language tasks including section identification,[20] requires significant high-quality data and computational resources. As real-world data drifts over time and relative model weights must be updated to reflect the changes, production solutions utilizing these static models also demand infrastructure and expense for substantial model re-training and deployment process overhead. Practically, this exacerbates the total cost of translating and operating continual-pretrained and fine-tuned models in clinical applications.

Prompting methods for transformer-based, instruction-tuned, and foundation-pretrained language models, on the other hand, have been shown to perform comparably and even exceed the performance of their fine-tuned model counterparts across a diversity of tasks without adhering rigidly to a single domain.[22–25] In medical applications, any tendency to hallucinate or "confabulate" is problematic with the low tolerance for error in medical applications, but advanced prompt engineering techniques have been shown to significantly mitigate these errors, especially on more complex tasks and data.[22] Between these techniques, few-shot prompting specifically has exhibited the most consistent and dramatic performance increases on clinical language tasks such as clinical sense disambiguation,[23] tabular text classification,[26] disease prediction,[27] medication extraction[23,28] and social determinants of health (SDoH) extraction,[29] among other biomedical and clinical applications.[23,30,31]

Few-shot prompting improves performance across language tasks by associating domain-specific input samples to their appropriate labels and providing tangible instances of the expected output format, allaying unintentional ambiguity in the provided class definitions and formatting instructions that are otherwise challenging to convey in zero-shot settings. As LLM models have been constrained by the number of exemplars able to fit in the input context window and the number of labelled samples often exceeds this window, dynamically selecting which exemplars to include based on similarity of the exemplar's question to the current input's question has demonstrated a significant, additional performance improvement over static or randomly selected examples,[32,33] including on clinical language tasks such as clinical text summarization,[34] medical dialogue summarization,[35] adverse drug event (ADE) extraction,[36] medical named entity recognition,[37] and in a chain-of-thought reasoning process for medical error detection and correction.[38] However, this method has not been applied to section identification tasks such as section classification.

Although proprietary- and domain-specific LLMs such as GPT4, Gemini and Med-PaLM have displayed the strongest performance on clinical NLP tasks,[39] the caveats of evolving real-world data, lack of publicly available and deidentified datasets, and high cost of training, inference and maintenance of these proprietary and/or supervised models make clinical application untenable for many less-endowed healthcare institutions. These concerns instead make smaller, open-source models expedient for clinical applications. A recent review on biomedical applications of LLMs reported a consensus among both open-source and proprietary LLM researchers that prompting techniques such as few-shot learning already significantly improve LLM performance and will continue to advance, but that much more research is needed into how best to induce the optimal responses and which LLM orchestration architectures are best, specifically with open-source and reasonably lightweight, lower-parameter models.[19]

In this study, we propose a dynamic few-shot prompting method for clinical note section classification using lightweight, open-source LLMs. We compare the performance of this method against zero-shot and randomly selected static shot scenarios using a suite of recently released open-source models. We also analyze the performance of a selected subset of these models with limited embedding samples and varying amounts of exemplars to assess the relationship between training set size and classification accuracy to ascertain the number of resources necessary for adequate results without exhaustive annotation and fine-tuning. Our aim is to present a practical solution for determining document structure from arbitrary clinical text to ultimately improve the performance of downstream tasks and end-to-end clinical applications.

Materials and methods

Data

Most previous section classification modeling efforts have opted to annotate and evaluate their models on private data, due in part to the high costs and risks that come with deidentification. As a result, a survey we conducted on clinical note section classification and document structure awareness prior to this study indicated a scarcity of publicly available datasets that can be used for section classification and that conform to standardized ontologies. Nonetheless, we identified a publicly available dataset based on MIMIC-III as well as a deidentified Mayo Clinic corpus to evaluate our models against the section classification task. Table 1 lists some attributes of each dataset and task, including the raw training and test set counts after splitting.

Downloaded from https://academic.oup.com/jamia/article/32/7/1164/8155977 by Mayo Clinic Library user on 30 June 2025

Table 1. Section classification task dataset average sample length, label, data split and counts for each dataset.

Task	Dataset Source	Schema	Sections	Labels	Input Length[a]	Training	Test
MPS	MIMIC-III	SOAP	4	Single	879.1	3566	595
MCS	Mayo	SOAP	4	Multiple	171.9	2826	706
MCC	Mayo	CDA[b]	15				

[a] Average number of characters per sample input.
[b] Modified CDA 1.0 schema includes 15 total sections.

MPS dataset

The first dataset served as the Progress Note Understanding task from the 2022 n2c2 shared task track 3,[40,41] encompassing 603 training and 87 test ICU progress notes derived from the MIMIC-III Note Events table data. The full MIMIC-III dataset contains over 40 000 ICU admission records including progress notes. The derived MIMIC progress note understanding (MPS) corpus used in our evaluation was annotated to include the SOAP section information. Each progress note sample contains a line-by-line annotation of Subjective, Objective, Assessment, or Plan (SOAP) section labels with the explicit section headers from the original document omitted for the shared challenge. We processed these notes by parsing and concatenating each contiguous group of lines labelled with the same SOAP identifier into separate sections from each provided train and test spreadsheet, creating 3566 training and 595 test section text samples.

Mayo clinic multi-EHR SOAP (MCS) and CDA (MCC) dataset

While the SOAP structure provides a degree of context within a clinical note, the lack of granularity misses nuances that subsection information might capture, such as whether the patient's social history or current medications are being discussed regarding a treatment mention. To leverage a more refined schema that makes patient data transferable across data sources and is compliant with their EHR web technology, many healthcare institutions have adopted the Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) standard or its predecessor, Clinical Document Architecture (CDA), as a clinical documentation practice.[42] Another dataset we include is a corpus of 6000 clinical note text segments of up to three contiguous sentences, labelled both by SOAP section and CDA Release 1 section. The clinical notes were collected across Mayo Clinic sites spanning three distinct electronic health records (EHRs), GE Centricity (GEC), Cerner, and Epic, over ten years. Two subject matter expert annotators and one adjudicator produced the labels and deidentified the samples. The Cohen Kappa inter-annotator agreement for the SOAP section annotation is 0.67, 0.83, and 0.99 for GEC, Cerner, and Epic, respectively, and 0.53, 0.73, and 0.84, respectively, for the CDA section labels. This dataset is currently pending for publication and public release.

Few-shot prompting

Few-shot learning is a machine learning model training practice where model training is limited to a few labelled samples. Few-shot prompting of LLMs, also referred to as "in-context" learning, incorporates training samples into the prompt of a pretrained foundation model not having otherwise undergone fine tuning. Though the model has scant information from which to infer in most few-shot cases, few-shot prompting has demonstrated remarkable improvements in performance by providing templates of expected output format as well as samples for target classes that aid in framing the class definitions.

Dynamic few-shot prompting

Our proposed prompt engineering approach to section classification utilizes dynamic few-shot example selection process. Few-shot learning, a type of in-context learning, leverages explicit instructions and question-answer examples. Which examples to include in the prompt for a given input section can be determined *dynamically*, where examples are selected based on similarity to the current input.

To transform input text into a vector embedding representing a point in latent vector space, a separate transformer encoder model tuned specifically for generating embeddings is typically used. For this purpose, we opted for bge-base-en-v1.5, the sentence transformer-based embedding model developed by Baidu and available in the HuggingFace repository. At the start of our experimentation, this model was posting among the top accuracy scores on the HuggingFace Massive Text Embedding Benchmark (MTEB) leaderboard. The examples are selected by cosine similarity of the generated embedding for the given input section text to the section text samples representing the nearest context embeddings of the training set in the vector store. During inference, the test inputs are converted using the same embedding model then the generated context embedding is compared to the nearest training sample context embeddings. Figure 1 illustrates the process of splitting the dataset, embedding of training data, and retrieving examples pertinent to the test input in the dynamic few-shot prompting scenario.

Due to the 2k and 4k token maximum on the smaller models, the number of few shot examples able to fit within the context window of the prompt is constrained, especially for the older models included in our model comparison experiments. Thus, to use a consistent number of shots across model comparison experiments, we limited all static and dynamic few-shot tests to 8 total exemplars.

Contextual prompting

With dynamically selected examples, we also introduce a prompting method that includes varying degrees of contextual information surrounding the input section. In our experiments, this supporting contextual information includes different amounts of text adjacent to each input section. The prompts designate this additional text as supplemental and ask the model only to classify the text of the target input section between delimiters. We then assess the impact these different levels of supporting contextual information have by comparing performance to our dynamic and static few shot prompting methods which use only the input section. Figure 2 displays the prompt template used for our few-shot prompting scenarios.

Downloaded from https://academic.oup.com/jamia/article/32/7/1164/8155977 by Mayo Clinic Library user on 30 June 2025

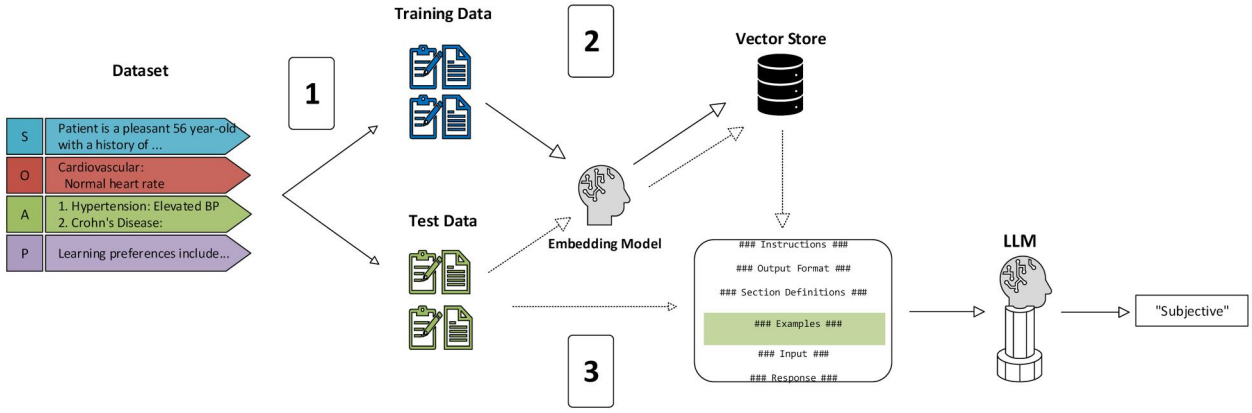

Figure 1. Dynamic few-shot prompting method. (1) Section text and labels are parsed from source corpus and split into training and test data. (2) Training data is embedded using embedding model and stored in vector embedding database. (3) During inference, contextual embeddings are generated from the given input text sample and the k most similar examples in vector store to the current input are retrieved, where k is the number of dynamic shots in the current experimental setting. The retrieved examples are then inserted into the Examples (green) section of the prompt and passed to the LLM for output generation.

```
### Instructions ###                                                    (a)
Your task is to classify the input text segment into one of the section categories

### Output Format ###                                                   (b)
Use the specified JSON Format: "{'section': '<category>'}"

### Section Definitions ###                                             (c)
- Subjective: Description of a patient's chief complaint, current symptoms, …
- Objective: Measurements such as vital signs, physical exam findings, …

### Examples ###                                                        (d)
Sample Input: Ms. Patient is a 33-year-old suffering from Crohn's …
Correct Response: {'section': 'Subjective'}

### Preceding Text ###                                                  (e)
Heart: 87 bpm …

### Input ###                                                           (f)
Mr. Patient is a pleasant 59-year-old male with visiting because of his…

### Succeeding Text ###                                                (g)
Continue with POC …

### Response ###                                                        (h)
```

Figure 2. Few-shot prompt template. The Instructions component (a) explains the system role as a clinical note assistant, informs the system that it will be given definitions and instructs to classify into one of the section categories. The Output Format component (b) specifies the structured JSON format that the output response should conform to for consistent parsing. The Section Definitions component (c) describes the purpose and information contained in each section category. The Examples component (d), only present in few-shot experimental prompts, contains several hardcoded and dynamically selected section text samples and their corresponding labels in the expected output format. The Preceding Text (e) and Succeeding Text (g) components, containing the text immediately before and after the current input section, respectively, and were only present in for Contextual Prompting inference. The Input component (f) contains the current section input context, while the Response heading (h) prompts the system for the label prediction.

Experimental models

We evaluate our section classification approach against seven open source LLMs: Llama 2 7B and 13B,[43] Llama 3.0 8B,[44] Mistral 7B v0.1,[45] Mistral 7B OpenOrca,[46] Mistral 7B v0.3,[47] and Gemma 2 9B.[48] Further information about these models can be found in the Models subsection of the Supplementary Materials. Details about each model's context window size and training methods are listed in Table S1.

Baseline settings

Zero-shot

We include a zero-shot setting where the Examples section is omitted from the prompt and only the task instructions, class definitions, and input text are given. This setting allows us to evaluate how each model performs with task in the absence of input-output pairs.

Static few-shot

Another baseline scenario uses few-shot exemplars that are randomly selected from the training datasets in lieu of the dynamically selected few-shot exemplars for our experimental scenario. Comparison of this static approach against our experimental dynamic selection should reveal the precise impact of the context-matching mechanism of dynamic selection for a given input to the embedded samples.

Experimentation

Four NVIDIA A100 GPUs, each with 40GB VRAM, were used to generate embeddings for training data in few-shot

Downloaded from https://academic.oup.com/jamia/article/32/7/1164/8155977 by Mayo Clinic Library user on 30 June 2025

settings and perform inference during evaluation. The temperature parameter was fixed at 0.0001, as Llama models do not allow a "deterministic" 0.0 temperature setting, and a consistent set of parameters across models was necessary for model comparison. *Top_k* sampling and *top_p* parameters were kept at the defaults of 50 and 1.0, respectively, for all experiments.

Evaluation

Model comparison

We calculate macro and micro F1 scores in each category as well as overall for every model. The class imbalance in the Mayo dataset, especially for model comparison on the CDA schema, necessitates the use of both micro and macro weighting of F1 for a more precise measurement of model performance across infrequent classes.

Ablation testing

To gauge the relationship between training set embedding size and performance, we conduct ablation testing in each few-shot scenario by limiting selected samples for each class in the vector store and compare performances across models. The number of samples for each class is fixed in each scenario and are randomly selected from the full training corpus prior to embedding.

Context window utilization

For models that both perform well in our model comparison and purport a larger context window capacity, we explore the effect of increasing amounts of dynamic few-shot exemplars on macro F1 performance. In this set of tests, we evaluate each of the selected models with 16, 24, 32, and 48 exemplars in the prompt.

Effect of contextual information

We evaluate our contextual prompting approach on our MPS dataset using three settings that incorporate additional context in the prompt: local, section, and full note. The local setting provides the concatenated five lines immediately prior to and following the target input section, the section setting includes both the section before and after the target, and the full note section inserts the entire note encompassing the target section. Each setting excludes all section headers, as they have been removed in the source dataset. We conduct a section-specific error analysis to ascertain whether the performance of each section might be differentially influenced by the provision of the contextual information.

Embedding similarity

We measure the similarity of pairs of embeddings for the entire training dataset, between retrieved examples and their given input training section text, and within and across samples of each SOAP section label for the MPS dataset. The mean cosine similarity of these embedding pair groups is calculated.

Results

The macro-averaged F1 score for each of the baseline and dynamic few-shot methods for each of the model comparison, ablation testing, context window utilization, and contextual prompting experiments are presented here, as well as the section-level performance of the few-shot and contextual prompting scenarios. Figure 3 illustrates the model comparison results for each of the zero-shot and static few-shot baseline settings as well as the experimental dynamic few-shot method. The complete results are listed in Table S2. Across all models for both SOAP classification tasks, the dynamic few-shot prompting method averaged a 39.3% increase and 21.1% increase over the macro F1 scores of the zero-shot and static few-shot prompting methods, respectively. Across all models of the CDA classification task, the dynamic few-shot method yielded a 17.7% increase and a 16.7% increase over the zero-shot and static-few shot methods, respectively.

Figure 4 illustrates the results of ablation testing, where decreasing-sized subsets of the training data were randomly selected for embedding. The macro F1 values used in this illustration are listed in Table S3 of Appendix.

Table 2 displays the results of the context window utilization experiments, where each column header number represents the number of exemplars included in the prompt. Mistral 7B OpenOrca achieved the highest macro F1 score with only a 16-shot prompt, while for Llama 3.0 8B and Mistral 7B v0.3, 24 exemplars yielded optimal results. None of the three models we tested performed best in the two test scenarios containing the most exemplars.

The mean cosine similarity of the embedded training data sample pairs for the MPS data set is 0.766. Table S4 contains the mean cosine similarity of all section text samples in the training dataset as well as the mean cosine similarity of 8 and 16 dynamically retrieved exemplar embeddings to each of the input SOAP sections in the MPS dataset. Table S5 contains the mean cosine similarity of all embedding pairs both within each section and across each pair of section classes.

Table 3 enumerates the performance of the three contextual prompting methods we evaluated, while Table 4 shows the section-level precision, recall, and macro-averaged F1 scores for each section as well as overall, for the best performing contextual performance run of the MPS dataset, which included the full note context using Llama 3.0 8B model.

Discussion

Dynamic few-shot evaluation

For all tasks and models, the prompts incorporating the dynamically selected few-shot exemplars outperformed all baselines, in terms of Macro F1 scores, including the prompts using the same number of randomly selected few-shot exemplars.

The Mayo Clinic dataset represents more diverse source data as the corpus was not limited to ICU progress notes but multiple medical specialties over different Mayo campuses and EHRs. In the MCC task, random selection of exemplars did not engender a significant improvement over zero-shot baseline in any model, even reducing the performance of Llama 2 7B, Llama 3.0 8B, and Gemma 2 9B. For Mistral 7B OpenOrca, which achieved the highest F1 scores for the MPS and MCC tasks (0.935 and 0.584, respectively), the difference between zero-shot and static few-shot was ostensibly insignificant. For every model and task, however, the increase in macro F1 from static to dynamic few-shot was stark, averaging 20.1 percent across all models. As each dataset represents a diversity of notes across a range of patient demographics, encounter types, and note-taking styles, the heterogeneity of input texts and number of labels in the classification task both appear to be associated with the

Downloaded from https://academic.oup.com/jamia/article/32/7/1164/8155977 by Mayo Clinic Library user on 30 June 2025

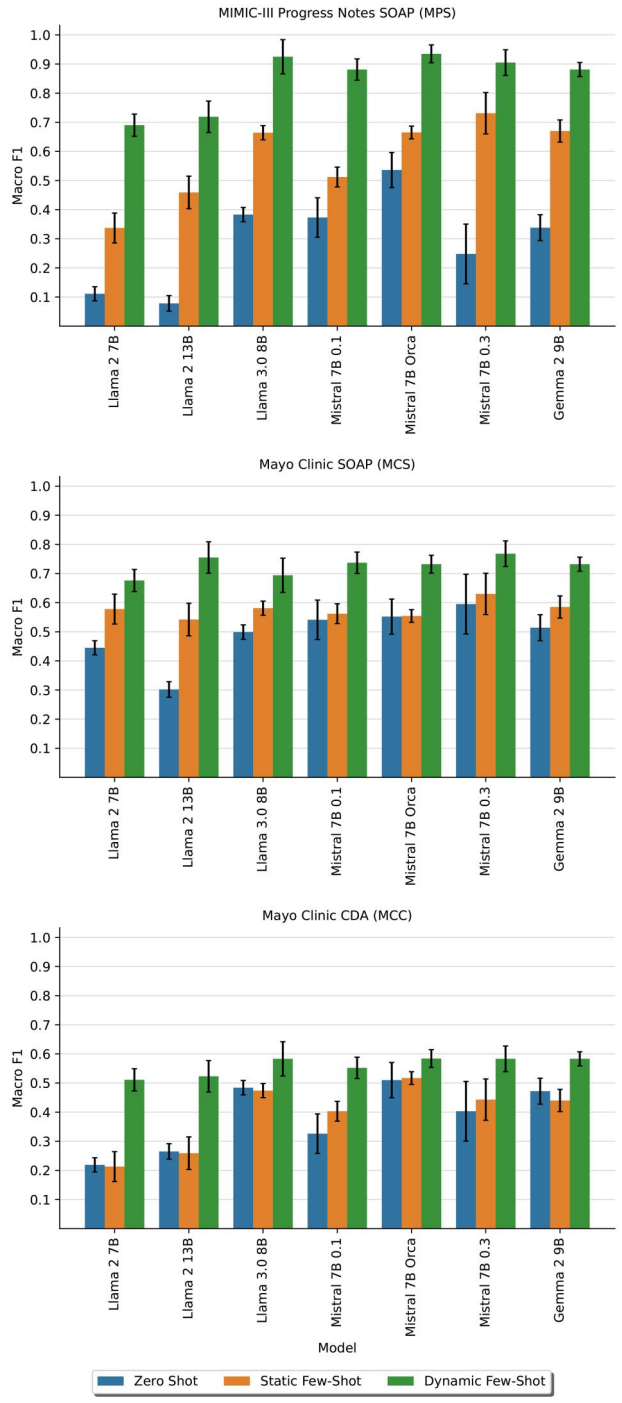

Figure 3. Model performance comparison. Each grouped bar chart displays the performance of the models across the baseline and experimental prompt settings for a section classification task. Bootstrapped 95% confidence intervals are represented in each bar.

remarkable advantages conferred by dynamic selection of few-shot exemplars.

Since the CDA classification dataset includes an imbalanced label distribution with several more classes overall including infrequent and rare classes, evaluation of our approach on this dataset is intended to showcase how the method performs in these commonly skewed real-world data distributions. For the CDA classification task in our model comparison, the lack of improvement from zero-shot to static few-shot yet dramatic improvement in macro F1 scores from static to dynamic few-shot for each model underscores the importance of matching similar examples in the rare class cases, since the macro F1 metric weighs rare classes disproportionately more than common classes.

Model performance across tasks

Across both SOAP tasks, the models appear to perform better in zero-shot and static few-shot in the Mayo dataset than on the labelled progress notes from the MIMIC-III dataset. However, the overall F1 scores achieved by the dynamic few-shot method are much higher in 5 out of the 7 models (all models except Llama 2 7B and 13B). The degree of difference in performance between each setting in the progress notes MIMIC-III dataset is also much greater, as the static few-shot was significantly better than zero-shot, and dynamic was significantly better than static for all models, whereas static few-shot was not significantly better than zero-shot in the Mayo Clinic SOAP task.

For the same Mayo Clinic dataset, performance across the two tasks also varied. In the CDA classification task, which includes a much greater number of classes and more class imbalance, the overall scores were lower for all models. The static few-shot scenario additionally conferred no significant increase in performance, whereas the dynamic few-shot scenario bore a significant increase for most models.

Model comparison

The inclusion of two sizes of Llama 2 models, 7B and 13B, facilitates an analysis of parameter space on the performance in the tasks. While the 13B-parameter version of Llama 2B performed better than 7B in all three tasks using the dynamic few-shot approach, 13B performed worse in all three tasks in the zero-shot approach. A possible reason for this is that the additional sparsity of the 13B model only elevated the ambiguity of terms in the class definitions. In the absence of input-output sample pairs, the sole reliance of the model on the class definitions increased the effect of Llama 2's pretraining data noise, thereby counterintuitively reducing the performance with additional parameters.

The addition of three versions of Mistral 7B models (Mistral 7B v0.1, Mistral 7B OpenOrca, and Mistral 7B v0.3) allows a relatively controlled evaluation of the effect of the different pre-training and tuning methods used by the contributors to improve those models. The instruction-tuned OpenOrca variant achieved the highest F1 score in the MPS and MCC tasks, despite its release just after v0.1 and preceding v0.3 by roughly 9 months. The advantage of instruction tuning demonstrated in our evaluation corroborates the findings in Zhang et al[49] for general NLP and in Tran et al[50] for biomedical NLP, where instruction-tuned models exhibited significant performance improvements over non-instruction-tuned baselines.

Ablation study

In the ablation tests, the restriction of embedding resources through random selection of subsets of the training data appears to generally reduce performance of the model, as expected. However, the effect is negligible or absent in certain cases. For example, Llama 3.0 8B performance does not decrease when training samples are limited from 1600 to 80 (0.913 vs 0.919 Macro F1, respectively). Llama 3.0 8B seems to perform comparatively well in the 40- and 80-embedding-

Downloaded from https://academic.oup.com/jamia/article/32/7/1164/8155977 by Mayo Clinic Library user on 30 June 2025

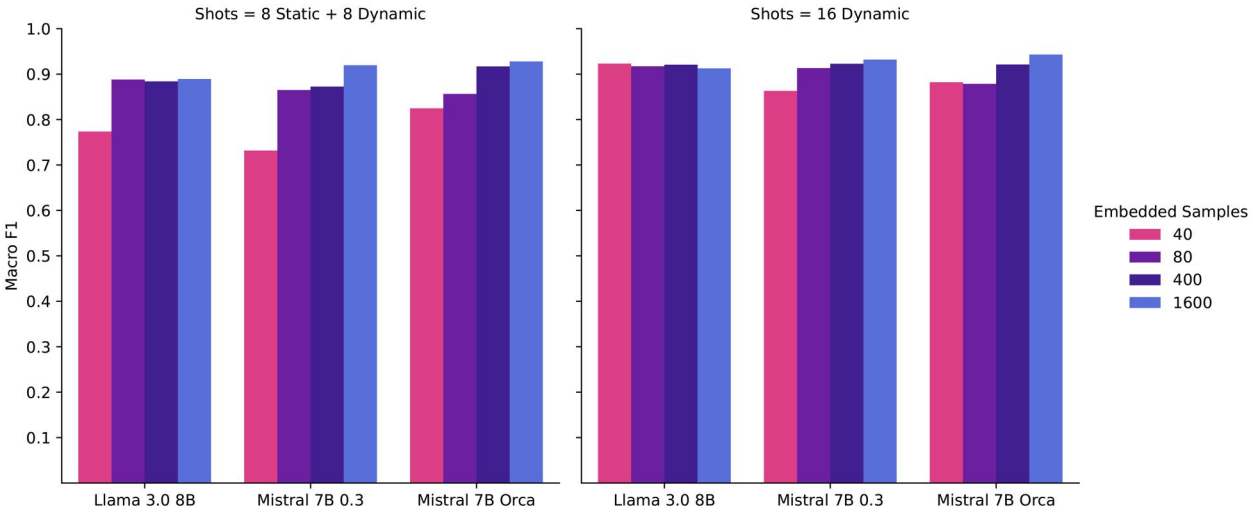

Figure 4. Ablation testing results. Each bar represents the macro F1 score for a different number of embedded training samples in the vector store.

Table 2. Context window utilization.[a]

Model	Exemplars			
	16	24	32	48
Llama 3.0 8B	0.938/0.947	**0.948/0.955**	0.687/0.851	0.513/0.674
Mistral 7B OO	**0.941/0.943**	0.731/0.925	0.611/0.929	0.692/0.876
Mistral 7B 0.3	0.927/0.937	**0.947/0.953**	0.941/0.945	0.945/0.952

Selected models' performance using different exemplars on the MPS dataset.
[a] Scores shown are Macro F1/Micro F1. The scores in bold indicate the highest Macro/Micro F1 scores for each model.

Table 3. Contextual prompting performance.

Model	Context Type		
	Local[a]	Section[b]	Full Note
Llama 3.0 8B	0.949	0.951	**0.953**
Mistral 7B OO	0.929	0.914	0.909
Mistral 7B 0.3	0.922	0.925	0.931

Selected model's macro F1 performance when including note context in the prompt. The bold value indicates the best macro F1 score across models and contextual prompting scenarios tested.
[a] The local context includes the 5 lines of note text immediately before and after the given input section.
[b] The section context includes the text of the section prior to and following the given input section.

sample scenarios as it achieves the highest scores for both pure dynamic few-shot and hybrid experiments.

Context utilization analysis

The best performance of each model, shown in bold in Table 2, indicates that all three of the models tested performed best with equal to or fewer than 24 exemplars. Though these models purport context windows that can accommodate many more exemplars, the clear degradation of performance for Llama 3.0 8B and Mistral 7B OpenOrca with additional exemplars exemplifies meager contextual information carryover from one sliding window to the next. In comparison, the relatively minor performance decrease of Mistral 7B v0.3 shows an advancement of sliding context window information carryover. However, the lack of a significant increase in performance as exemplars are added for this model reveals that there is a limit to the performance increase conferred by dynamically selecting examples, and that for our section classification task, the optimal number of samples for each of the tested models is around 24.

Effect of contextual information

The contextual prompting results displayed in Table 3 showcase the performance improvement imparted by commuting additional contextual information in the prompt of the relative position of the section within the note and the adjacent note text, for 2 out of the 3 tested models. Although the increase in F1 score may not be substantial enough to warrant the additional cost of prompt tokens, the results show that the inclusion of contextual information nonetheless appears to render higher accuracy. The lack of improvement for Mistral 7B OpenOrca, despite its superior performance in our model comparison tests, indicates that its instruction tuning and inferior context window capacity inhibited the utility of the additional context information, as the context requires more context breadth and more instruction than the prompts without such contextual information.

Training data and section similarity

The mean cosine similarities of all embedded training data section text sample pairs and of each pair of retrieved examples to the current input section text, in 8-shot and 16-shot exemplar settings, is shown in Table S4. The higher similarity of 8-shot retrieved exemplars to the current input text than the 16-shot retrieved exemplars to the current input text seems appropriate as the additional less-similar exemplars

Downloaded from https://academic.oup.com/jamia/article/32/7/1164/8155977 by Mayo Clinic Library user on 30 June 2025

Table 4. Performance by section.

Section	Static Few-Shot			Dynamic Few-Shot			Contextual Dynamic Few-Shot		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Subjective	0.934	0.934	0.934	0.963	0.917	**0.939**	0.957	0.917	0.936
Objective	0.874	0.988	0.928	0.988	0.994	0.991	1.0	0.997	**0.997**
Assessment	0.635	0.930	0.755	0.975	0.907	0.940	0.988	0.942	**0.964**
Plan	0.958	0.267	0.418	0.828	0.953	0.886	0.857	0.977	**0.913**
Overall	0.850	0.780	0.759	0.938	0.924	0.939	0.950	0.957	**0.953**

The precision, recall, and macro F1 score for each SOAP section label of the MPS dataset using Llama 3.0 8B with static few-shot, dynamic few-shot, and contextual dynamic few-shot prompting with 16 shots.
The highest F1 score for each section label and overall is shown in bold.

would increase the mean cosine distance. In Table S5, The cosine similarities shown along the diagonal represent the cohesiveness of examples within each class, whereas the remaining values represent the mean cosine similarity (inverse mean cosine distance) between all samples of one class with each counterpart class. The classes with the highest mean cosine similarity of embedding pairs are the Subjective and Assessment classes (0.770), whereas the Subjective and Plan had the lowest mean similarity of any pair (0.705). This confirms the hypothesis that Subjective and Assessment sections are most similar since they often contain the same first few sentences introducing the patient and describing their primary symptoms or concerns. The often repetitive, templated language in Assessment sections also explains the higher cohesion of samples within the Assessment class (0.853).

Performance analysis by section

As exhibited in Table 4, Llama 3.0 8B achieved the highest overall macro F1 score (0.953) with the inclusion of 16 dynamic few-shot exemplars and contextual information in the prompt, where the contextual information provided was the full progress note, including all text prior to and following the current input section. The section-specific performance reveals high performance on inference of Objective sections, yet a substantially underperforming Plan section. This was unexpected since the language in Plan sections is often prescriptive, including recommendations for future testing, medication, or surgical treatments to prescribe as opposed to those historically taken or undergone, referrals, and patient education, which are all relatively unique among the Plan sections. The concordance of Plan section as the lowest F1 score for each type of few-shot prompting could indicate a lack of sufficient differentiation of future-tense verb embeddings from their root token embeddings, as well as other semantic information unique to the Plan section that would otherwise hypothetically signal the identity of the Plan section to the model.

Limitations

Since the scope of this study was constrained to practical models that are probable for incorporation and translation into clinical applications, we did not include many proprietary models that lead in performance on general NLP benchmarks such as GPT-4, Claude, and Gemini. Though these models may achieve marginally better performance on clinical and non-clinical tasks than open-source models with fewer parameters, the burdens of cost and privacy make these closed-source models prohibitive in most clinical settings and were therefore not included in our comparison.

While the SOAP note-taking schema is employed widely across institutions, many note types included in the Mayo dataset, such as correspondence letters and imaging reports, are not written in the SOAP (or other section schema) template. Additionally, the definitions of each section conferred to the annotators of this dataset differed from those of the MPS dataset, hence multiple possible labels were applied to each text segment. This limitation precluded cross-domain evaluation, where one SOAP dataset was used for training while the other was used for testing. The assignment of multiple labels to each input sample in the Mayo dataset also prevented comparison to sentence transformer baseline models, as the fine-tuning of transformer models necessitates either a single class (multi-class task) or multiple required labels (multi-label task), but not multiple optional labels.

Future work

The results of this project suggest several potential avenues for future research. First, the remarkable performance of the instruction-tuned Mistral 7B OpenOrca model on each task warrants further instruction-tuned model comparison as well as custom instruction-tuning on specific clinical language understanding tasks. Since we used a sentence-transformer embedding model and compact, robust LLM-based embedding models are now available, an investigation of these more advanced models containing larger context windows on their impact in a dynamic example selection strategy is advisable.

The heterogeneity of clinical note structure and semantics across healthcare institutions also suggests a demand for cross-domain analysis to evaluate the portability of this dynamic example retrieval technique without the need for institution-specific datasets.

The superiority of Llama 3.0 8B model in the contextual prompting evaluation begs the question of whether more recent LLMs purporting even larger and more efficient context windows will even more effectively leverage the contextual information presented in the prompt. Further evaluation could be conducted with other types of contextual information, such as clinical note metadata, input section position information relative to the total number of note sections, and/or abstractive note summaries prior to section classification inference.

Finally, subsequent research could include analysis into how section classification, as well as other document structure and metadata extraction strategies, influence the performance of downstream tasks and applications such cohort retrieval and clinical trial matching systems by adding contextual information.

Downloaded from https://academic.oup.com/jamia/article/32/7/1164/8155977 by Mayo Clinic Library user on 30 June 2025

Conclusion

In this methodology study, we explore a dynamic few-shot prompting approach to a clinical note section classification task and evaluate the performance of lightweight, open-source models. In doing so, we assess the viability of LLMs for extracting clinical document structure for downstream practical applications. We compare the dynamic example selection approach to static few-shot and zero-shot baseline scenario performances. The results demonstrate the superiority of the dynamic few-shot method without the need for large, labelled corpora and suggest a need for further research into similar methods for use in other clinical language systems.

Author contributions

Kurt Miller contributed to the conception and design of the work, implementation of the methods, acquisition of the data, analysis and interpretation of the data, and drafting and editing the manuscript. Stephen Bedrick contributed to the design of the work, proposed the analysis, created visualizations for the analysis, and drafting of the manuscript. Qiuhao Lu contributed to the data analysis, interpretation of the data, and drafting and editing the manuscript. Andrew Wen contributed to the design of the work, interpretation of the data, and reviewing the manuscript. William Hersh contributed to the design of the work, interpretation of the data, and reviewing the manuscript. Kirk Roberts contributed to the conception and review of the work. Hongfang Liu contributed to the conception and design of the study, acquisition of the data, analysis and interpretation of the data, and reviewing each version of the manuscript.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by the National Institutes of Health grant number R01LM011934.

Conflicts of interest

None declared.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

References

1. Zullig LL, Whitson HE, Hastings SN, et al. A systematic review of conceptual frameworks of medical complexity and new model development. *J Gen Internal Med*. 2016;31:329-337. https://doi.org/10.1007/s11606-015-3512-2
2. Edwards ST, Neri PM, Volk LA, Schiff GD, Bates DW. Association of note quality and quality of care: a cross-sectional study. *BMJ Qual Saf*. 2014;23:406-413. https://doi.org/10.1136/bmjqs-2013-002194
3. Fu S, Leung LY, Raulli A-O, et al. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. *BMC Med Inf Decis Making*. 2020;20:60. https://doi.org/10.1186/s12911-020-1072-9
4. Rule A, Bedrick S, Chiang MF, Hribar MR. Length and redundancy of outpatient progress notes across a decade at an academic medical center. *JAMA Netw Open*. 2021;4:e2115334. https://doi.org/10.1001/jamanetworkopen.2021.15334
5. Searle T, Ibrahim Z, Teo J, Dobson R. Estimating redundancy in clinical text. *J Biomed Inf*. 2021;124:103938. https://doi.org/https://doi.org/10.1016/j.jbi.2021.103938.
6. Fu S, Vassilaki M, Ibrahim OA, et al. Quality assessment of functional status documentation in EHRs across different healthcare institutions. *Front Digital Health*. 2022;4:958539. https://doi.org/10.3389/fdgth.2022.958539 [published Online First: 2022/10/15].
7. Miller K, Moon S, Fu S, Liu H. Contextual variation of clinical notes induced by EHR migration. *AMIA Annu Symp Proc*. 2023;2023:1155-1164. [published Online First: 2024/01/15].
8. Mashima Y, Tanigawa M, Yokoi H. Information heterogeneity between progress notes by physicians and nurses for inpatients with digestive system diseases. *Sci Rep*. 2024;14:7656. https://doi.org/10.1038/s41598-024-56324-7
9. Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. A comprehensive study of named entity recognition in Chinese clinical text. *J Am Med Inf Assoc*. 2014;21:808-814. https://doi.org/10.1136/amiajnl-2013-002381 [published Online First: 2013/12/19].
10. Edinger T, Demner-Fushman D, Cohen AM, Bedrick S, Hersh W. Evaluation of clinical text segmentation to facilitate cohort retrieval. *AMIA Annu Symp Proc*. 2017;2017:660-669. [published Online First: 2018/06/02].
11. Richter-Pechanski P, Geis NA, Kiriakou C, Schwab DM, Dieterich C. Automatic extraction of 12 cardiovascular concepts from German discharge letters using pre-trained language models. *Digital Health*. 2021;7:20552076211057662. https://doi.org/10.1177/20552076211057662 [published Online First: 2021/12/07].
12. Doan S, Bastarache L, Klimkowski S, Denny JC, Xu H. Integrating existing natural language processing tools for medication extraction from discharge summaries. *J Am Med Inf Assoc*. 2010;17:528-531. https://doi.org/10.1136/jamia.2010.003855 [published Online First: 2010/09/08].
13. A Supervised Abbreviation Resolution System for Medical Text. Conference and Labs of the Evaluation Forum; 2013.
14. Dolin RH, Alschuler L, Beebe C, et al. The HL7 clinical document architecture. *J Am Med Inf Assoc*. 2001;8:552-569. https://doi.org/10.1136/jamia.2001.0080552 [published Online First: 2001/11/01].
15. Patterson O, Hurdle JF. Document clustering of clinical narratives: a systematic study of clinical sublanguages. *AMIA Annu Symp Proc*. 2011;2011:1099-1107. [published Online First: 2011/12/24].
16. Doing-Harris K, Patterson O, Igo S, Hurdle J. Document sublanguage clustering to detect medical specialty in cross-institutional clinical texts. *Proc ACM Int Workshop Data Text Min Biomed Inform*. 2013;2013:9-12. https://doi.org/10.1145/2512089.2512101 [published Online First: 2013/10/01].
17. Sohn S, Wang Y, Wi C-I, et al. Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. *J Am Med Inf Assoc*. 2018;25:353-359. https://doi.org/10.1093/jamia/ocx138
18. Pomares-Quimbaya A, Kreuzthaler M, Schulz S. Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. *BMC Med Res Methodol*. 2019;19:155. https://doi.org/10.1186/s12874-019-0792-y
19. Sahoo SS, Plasek JM, Xu H, et al. Large language models for biomedicine: foundations, opportunities, challenges, and best practices. *J Am Med Inf Assoc*. 2024;31:2114-2124. https://doi.org/10.1093/jamia/ocae074
20. Zhou W, Yetisgen M, Afshar M, Gao Y, Savova G, Miller TA. Improving model transferability for clinical note section

Downloaded from https://academic.oup.com/jamia/article/32/7/1164/8155977 by Mayo Clinic Library user on 30 June 2025

classification models using continued pretraining. *J Am Med Inf Assoc*. 2023;31:89-97. https://doi.org/10.1093/jamia/ocad190
21. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv*. 2023;55:1-35. Article 195. https://doi.org/10.1145/3560815
22. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Med Inf*. 2024;12:e55318. https://doi.org/10.2196/55318
23. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large Language Models are Few-Shot Clinical Information Extractors. Accessed May 01, 2022. https://ui.adsabs.harvard.edu/abs/2022arXiv220512689A.
24. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. Accessed May 01, 2020. https://ui.adsabs.harvard.edu/abs/2020arXiv200514165B.
25. Taylor N, Zhang Y, Joyce DW, Gao Z, Kormilitzin A, Nevado-Holgado A. Clinical prompt learning with frozen language models. *IEEE Trans Neural Networks Learn Syst*. 2024;35:16453-16463. https://doi.org/10.1109/TNNLS.2023.3294633
26. Hegselmann S, Buendia A, Lang H, Agrawal M, Jiang X, Sontag D. TabLLM: few-shot classification of tabular data with large language models. Accessed October 01, 2022. https://ui.adsabs.harvard.edu/abs/2022arXiv221010723H.
27. Cui H, Shen Z, Zhang J, et al. 2024. LLMs-based few-shot disease predictions using EHR: a novel approach combining predictive agent reasoning and critical agent instruction. arXiv:2403.15464. preprint: not peer reviewed.
28. Gero Z, Singh C, Cheng H, et al. 2023. Self-verification improves few-shot clinical information extraction. arXiv:2306.00024.
29. Guevara M, Chen S, Thomas S, et al. Large language models to identify social determinants of health in electronic health records. *npj Digital Med*. 2024;7:6. https://doi.org/10.1038/s41746-023-00970-0
30. Ge Y, Guo Y, Das S, Al-Garadi MA, Sarker A. Few-shot learning for medical text: a review of advances, trends, and opportunities. *J Biomed Inf*. 2023;144:104458. https://doi.org/https://doi.org/10.1016/j.jbi.2023.104458.
31. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172-180. https://doi.org/10.1038/s41586-023-06291-2
32. What Makes Good In-Context Examples for GPT-3?; 2022 May; Dublin, Ireland and Online. Association for Computational Linguistics.
33. D'Abramo J, Zugarini A, Torroni P. Dynamic Few-Shot Learning for Knowledge Graph Question Answering. Accessed July 01, 2024. https://ui.adsabs.harvard.edu/abs/2024arXiv240701409D.
34. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med*. 2024;30:1134-1142.
35. SummQA at MEDIQA-Chat 2023: In-Context Learning with GPT-4 for Medical Summarization. Clinical Natural Language Processing Workshop; 2023.
36. Berkowitz J, Srinivasan A, Cortina JMA, Tatonetti1 N. TLab at #SMM4H 2024: retrieval-augmented generation for ADE extraction and normalization. Proceedings of the 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks 2024.
37. Li M, Zhou H, Yang H, Zhang R. RT: a retrieving and Chain-of-Thought framework for few-shot medical named entity recognition. *J Am Med Inf Assoc*. 2024;31:1929-1938. https://doi.org/10.1093/jamia/ocae095
38. KnowLab_AIMed at MEDIQA-CORR 2024: Chain-of-Though (CoT) prompting strategies for medical error detection and correction; 2024 June; Mexico City, Mexico. Association for Computational Linguistics.
39. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29:1930-1940. https://doi.org/10.1038/s41591-023-02448-8
40. Gao Y, Dligach D, Miller T, et al. Hierarchical annotation for building a suite of clinical natural language processing tasks: progress note understanding. *LREC Int Conf Lang Resour Eval*. 2022;2022:5484-5493. [published Online First: 2022/08/09].
41. Gao YCJ, Miller T, Sharma B, Churpek M, Dligach D, Afshar M. Tasks 1 and 3 from Progress Note Understanding Suite of Tasks: SOAP Note Tagging and Problem List Summarization (version 1.0.0). 1.0.0 ed: PhysioNet.
42. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The fast health interoperability resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. *JMIR Med Inf*. 2021;9:e21929. https://doi.org/10.2196/21929
43. Touvron H, Martin L, Stone K, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. Accessed July 01, 2023. https://ui.adsabs.harvard.edu/abs/2023arXiv230709288T.
44. Dubey A, Jauhri A, Pandey A, et al. The Llama 3 Herd of Models. Accessed July 01, 2024. https://ui.adsabs.harvard.edu/abs/2024arXiv240721783D.
45. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. Accessed October 01, 2023. https://ui.adsabs.harvard.edu/abs/2023arXiv231006825J.
46. Bleys LWG, Pentland E, Cook A, Vong C. "Teknium". Mistral-lOrca: Mistral-7B Model Instruct-tuned on Filtered OpenOrcaV1 GPT-4 Dataset. https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca: Huggingface, 2023.
47. Mistral-7B-v0.3 repository on HuggingFace. Secondary Mistral-7B-v0.3 repository on HuggingFace 2024. https://huggingface.co/mistralai/Mistral-7B-v0.3.
48. Team G, Riviere M, Pathak S, et al. Gemma 2: Improving Open Language Models at a Practical Size. Accessed July 01, 2024. https://ui.adsabs.harvard.edu/abs/2024arXiv240800118G.
49. Zhang S, Dong L, Li X, et al. Instruction Tuning for Large Language Models: A Survey. Accessed August 01, 2023. https://ui.adsabs.harvard.edu/abs/2023arXiv230810792Z.
50. Tran H, Yang Z, Yao Z, Yu H. BioInstruct: instruction tuning of large language models for biomedical natural language processing. *J Am Med Inf Assoc*. 2024;31:1821-1832. https://doi.org/10.1093/jamia/ocae122
51. Bleys LWG, Pentland E, Cook A, Vong C. "Teknium". OpenOrca: An Open Dataset of GPT Augmented FLAN Reasoning Traces. August 2, 2023 ed. Huggingface repository: Huggingface, 2023.

© The Author(s) 2025. Published by Oxford University Press on behalf of the American Medical Informatics Association. All rights reserved.
For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.
Journal of the American Medical Informatics Association, 2025, 32, 1164–1173
https://doi.org/10.1093/jamia/ocaf084
Research and Applications

Downloaded from https://academic.oup.com/jamia/article/32/7/1164/8155977 by Mayo Clinic Library user on 30 June 2025