MoMA: A Mixture-of-Multimodal-Agents Architecture for Enhancing Clinical Prediction Modelling

Jifan Gao[1], Mahmudur Rahman[1], John Caskey[1], Madeline Oguss[1], Ann O'Rourke[1], Randy Brown[1], Anne Stey[2], Anoop Mayampurath[1], Matthew M. Churpek[1], Guanhua Chen[1,*], and Majid Afshar[1,*]

[1]University of Wisconsin-Madison, Madison, Wisconsin, USA
[2]Northwestern University, Chicago, Illinois, USA
[*]Corresponding authors: gchen25@wisc.edu, mafshar@medicine.wisc.edu

Abstract

Multimodal electronic health record (EHR) data provide richer, complementary insights into patient health compared to single-modality data. However, effectively integrating diverse data modalities for clinical prediction modeling remains challenging due to the substantial data requirements. We introduce a novel architecture, Mixture-of-Multimodal-Agents (MoMA), designed to leverage multiple large language model (LLM) agents for clinical prediction tasks using multimodal EHR data. MoMA employs specialized LLM agents ("specialist agents") to convert non-textual modalities, such as medical images and laboratory results, into structured textual summaries. These summaries, together with clinical notes, are combined by another LLM ("aggregator agent") to generate a unified multimodal summary, which is then used by a third LLM ("predictor agent") to produce clinical predictions. Evaluating MoMA on three prediction tasks using real-world datasets with different modality combinations and prediction settings, MoMA outperforms current state-of-the-art methods, highlighting its enhanced accuracy and flexibility across various tasks.

Keywords Multimodal learning · Large language models · Multi-agent system · Clinical predictions

Introduction

Modern healthcare increasingly leverages electronic health records (EHRs), which integrate diverse patient data modalities, such as clinical notes, medical images, vital signs, and laboratory results [1]. Each modality contributes unique, complementary information: clinical notes summarize patient symptoms, diagnoses, and treatments as documented by healthcare professionals; medical images objectively depict anatomical structures and pathology, facilitating disease detection and monitoring; laboratory and vital sign data quantify physiological states and abnormalities. The integration of multimodal EHR data enables a more holistic understanding of a patient's health conditions. The adaptation of multimodal EHR into a machine learning pipeline has been demonstrated to outperform those that only leverage a single modality in a wide range of clinical prediction tasks [2, 3, 4, 5, 6].

Multimodal integration methodologies typically fall into three categories: early fusion (concatenating inputs before training), joint fusion (co-learning representations during training), and late fusion (combining outputs from separately trained models) [7, 8]. Joint fusion is particularly promising, as it facilitates co-learning of a shared vector space, effectively capturing intricate cross-modal relationships. Various methods have been developed for learning this shared vector space, including cross-attention mechanisms [9], mixture-of-expert framework [10], contrastive learning [11], masked vision/language modeling [12], and variational approaches [13]. This method has consistently demonstrated superior performance compared to early and late fusion strategies across diverse clinical applications [14, 15, 16].

In particular, the recent surge of multimodal large language models (LLMs) is advancing the development of joint fusion methods. Vision language models, such as BLIP-2 [17], Flamingo [18], Kosmos-2 [19], and PaLM-E [20], have achieved state-of-the-art results while learning unified representations for images (or videos) and text from paired corpora. Beyond vision and text, ImageBind [21] and OneLLM [22] expand the shared space to six or more modalities, delivering competitive performance by contrastively aligning each additional signal to an image- or language-anchored representation. In the medical domain, LLaVA-Med [23], VILA-M3 [24], and GSCo [25] align radiology images with paired reports and achieved strong performance in various tasks.

All of these approaches, however, still depend on large paired multimodal datasets to learn a joint vector space, requiring non-trivial supervised alignment when a new modality is introduced incrementally. In healthcare, obtaining sufficient paired data is challenging due to the complexities associated with linking distinct modality-specific resources to the same patient or clinical encounter [26], coupled with inherent data fragmentation in healthcare systems [27]. These limitations present substantial obstacles to developing accurate multimodal models when lacking sufficient paired data for learning the shared vector space across various modalities.

Given these limitations, the fundamental motivation behind our method is to leverage the inherent capability of pretrained LLMs to translate multimodal clinical data into natural language. LLMs are widely recognized for their capability to capture semantic meaning from text [28, 29], including clinical notes [30]. When provided with clinical text inputs, LLMs can also serve as classifiers in clinical prediction tasks and have achieved outstanding performance [31, 32, 33]. Modern multimodal LLMs can also understand modalities beyond plain text, including medical images [23, 34] and structured EHR data [35, 36], and can convert non-plain text data into text summaries. Because of these rapid advances in multimodal LLMs, which demonstrate that rich semantics from non-text sources can be effectively translated into natural language, and the view from cognitive theory that language is central to human cognition [37], the converted text can serve as an aligned space, analogous to the shared vector space used in traditional joint-fusion multimodal approaches. Importantly, this conversion can be performed in a zero-shot manner using pretrained LLMs, avoiding the extensive data requirements typically associated with constructing such vector spaces.

In addition, studies have introduced the collaborative potential of LLMs whereby an LLM agent produces an improved response when incorporating outputs from other LLM agents [38]. This discovery suggests that introducing additional LLM agents, following the multimodal LLMs, to integrate original clinical text with LLM-generated summaries from non-text modalities could benefit predictive performance. Motivated by this observation, we propose the Mixture-of-Multimodal-Agents (MoMA) architecture for clinical prediction with multimodal EHR data (Figure 1). In MoMA, each non-text modality is processed by a pretrained, modality-specialized LLM agent that converts the non-text data into a corresponding text summary. These generated text summaries, along with existing clinical notes, are then integrated by an aggregator agent to form a unified narrative, which is used by a final predictor agent to generate clinical predictions. Details of the MoMA architecture are described in the Methods section.

The sequential transfer of these summaries harnesses the collaborative potential of LLMs, facilitating an effective integration of multimodal EHR data. MoMA can immediately leverage advances in any multimodal LLMs without retraining, because the non-text to text conversion can be swapped without retraining the rest of the system. Unlike existing multimodal LLMs that require paired multimodal data to learn a shared vector space, MoMA is an architecture that incorporates existing state-of-the-art multimodal LLMs in a plug-and-play fashion. Moreover, this architecture reduces training requirements by allowing the specialist and aggregator agents to operate in a zero-shot manner, with only the predictor agent requiring fine-tuning.

We validated the MoMA architecture on real-world, private datasets across three clinical tasks (chest trauma severity stratification, multitask chest and spine trauma severity stratification, and unhealthy alcohol use screening) with different combinations of modalities. Our results show that MoMA not only achieves superior performance in the overall testing set but also outperforms baseline models in every sex and race subgroup. Our ablation study reveals that the performance improvements benefit not only from the LLMs' text understanding capabilities but also from the effective integration of non-text modalities.

Traditional approaches to developing multimodal prediction models in healthcare require large volumes of high-quality paired EHR data to learn effective joint fusion representations, a requirement that is often unmet due to data quality challenges and strict privacy regulations that complicate the sharing of pretraining models [39, 40]. By leveraging open-source LLMs, our MoMA architecture circumvents these obstacles by translating non-text modalities into the natural language space without the need for extensive paired datasets. This not only reduces the resource burden associated with traditional joint fusion methods but also enables institutions with limited access to comprehensive multimodal data to develop accurate clinical prediction models.

Results

Datasets and Cohort Characteristics

We validated the MoMA architecture on three clinical prediction tasks using private datasets collected from the University of Wisconsin Hospitals and Clinics (UW Health): chest trauma severity stratification, multitask chest and spine trauma severity stratification, and unhealthy alcohol use screening. These tasks differ in complexity and classification structure: the first involves multiclass classification for chest trauma, the second jointly predicts multiclass severity for both chest and spine, and the third addresses binary classification for unhealthy alcohol use. They also involve distinct modality combinations: the first two tasks integrate free text clinical notes and chest radiographs, while the unhealthy alcohol use screening task combines free text and lab measurements.

Traumatic injuries are the leading cause of death among people younger than 45 [41]. Chest trauma is one of the most commonly encountered trauma injuries. Nearly half of trauma-related deaths occur after hospital admission [42] and timely stratification of chest trauma injury severity can help triage patients and predict complications [43]. The cohort used for the chest trauma severity stratification task was collected between January 2015 and December 2019, with a total sample size of 2,722 unique patients. This task involved a three-class classification of injury severity, annotated as negative, minor/moderate, and serious or greater. A team of certified trauma registrar coders at UW Health conducted extensive manual chart abstraction for each patient encounter, adhering to American College of Surgeons (ACS) and Trauma Quality Improvement Program (TQIP) standards [44] to calculate and validate Abbreviated Injury Scale (AIS) scores and associated trauma metrics [45]. Each patient encounter includes clinical notes as the text modality and chest radiographs as the non-text modality from the EHR.

Using the same cohort from the chest trauma severity stratification task, we defined a more complex multitask setting where models simultaneously predict injury severity for both the chest and spine. Each encounter is labeled using the same annotation protocol as the chest trauma severity stratification task, with injury severity assessed for the chest and spine. Each sample contains clinical notes as the text modality and chest radiographs as the non-text modality.

We also validated MoMA in an unhealthy alcohol use screening task. Alcohol misuse is recognized by the World Health Organization as one of the top five risk factors contributing to disease burden [46] and timely screening for unhealthy alcohol use can help mitigate the risk of alcohol-related harm [47]. In a prospective study, two research teams were deployed to screen, consent, and enroll 2,096 consented patients between September 2021 and February 2024, into the Tobacco, Alcohol, Prescription Medications, and other Substance (TAPS) screening tool, recommended by the National Institute on Drug Abuse [48], to assess unhealthy alcohol use in the past three months. This task uses whether a patient had unhealthy alcohol use in the previous three months as labels for binary classification. The emergency department (ED) recruitment team, comprising trained coordinators, screened willing patients for eligibility. Upon admission, an addiction medicine research team approached eligible patients to obtain informed consent and administer the TAPS screening tool, with participants receiving a gift card upon completion. The manual screen results were collected in a survey database and linked to the patient's related EHR encounter containing clinical notes as the text modality and lab measurements in tabular format as the non-text modality from the EHR.

We conducted temporal validation on both tasks to ensure the test sets are independent of the development set. For chest trauma severity stratification and multitask chest and spine trauma severity stratification, the development set used data collected from January 2015 to December 2018 and the test set used data collected from January 2019 to December 2019. For unhealthy alcohol use screening, the development set used data collected from September 2021 to August 2023 and the test set used data collected from September 2023 to January 2024. The characteristics of the two cohorts are shown in Table 1.

Overall Performance

For the chest trauma severity stratification task and the multitask chest and spine trauma stratification, we fine-tuned ClinicalBERT [49] with free text, following the methodology described in Gao et al. [5] as a published baseline. For unhealthy alcohol use screens, we compared our methods to a trained 1-dimensional convolutional neural network (1D-CNN) model from Afshar et al. [50], which processes clinical text mapped to medical concepts from the National Library of Medicine (CUI; concept unique identifier). These published baselines represent the current state-of-the-art (SOTA) approach for the three tasks (see the Methods section for more detailed descriptions). We also compared MoMA to LLaVA-Med [23], a widely used and representative multimodal LLM baseline in the medical domain [51, 52, 53], on both the chest trauma severity stratification task and the multitask chest and spine trauma stratification task. LLaVA-Med models were fine-tuned using the development set, with details described in the Methods section. Note that LLaVA-Med is not applicable to the unhealthy alcohol use screening task, as it does not support tabular data inputs. In addition, we evaluated the MoMA approach against two vector-based multimodal fusion methods: one using a cross-attention module to integrate representation vectors from two different modalities[54], and the other using a Mixture-of-Experts (MoE) mechanism for multimodal fusion[55]. Details of these two approaches are described in the Methods section. For the cross-attention and MoE baselines of both tasks and the published baseline of the chest trauma severity stratification and the multitask chest and spine trauma severity stratification task, we used the development and test sets as described in Table 1 to reimplement these methods. While we had access to the trained 1D-CNN model for unhealthy alcohol use screen positives, we did not have access to the original training dataset due to data restrictions. Therefore, the published baseline model had the advantage of being trained on a much larger dataset with a total of 54,915 encounters.

Since macro- and micro-F1 scores are standard metrics for multiclass classification [56, 57], we used them to evaluate chest trauma severity stratification and multitask chest and spine trauma stratification. For the binary task of unhealthy alcohol use screening, we used AUROC (Area Under the Receiver Operating Characteristic Curve) and AUPR (Area Under the Precision-Recall Curve), which are widely used in binary classification tasks [58, 59].

The performance metrics are summarized in Figure 2, where best baseline results, LLaVA-Med for chest trauma and multitask trauma stratification, and the published baseline for the unhealthy alcohol use screening, are highlighted using dotted lines for direct comparison; corresponding numerical results are provided in Supplementary Note 1. On both the chest trauma and multitask trauma stratification tasks, the MoMA outperformed all baselines, including the fine-tuned LLaVA-Med. Notably, it greatly outperformed the published baselines as well as the cross-attention and MoE baselines. In particular, for chest trauma severity stratification, both as a single task and within the multitask setting, MoMA achieved macro-F1 scores near 0.85 and micro-F1 scores above 0.90. For spine trauma stratification in the multitask setting, it achieved macro-F1 scores above 0.75 and micro-F1 scores close to 0.90. While a third modality (lab measurement) to MoMA was feasible for the trauma severity stratification tasks, we found that performance had already saturated with two modalities, and adding tabular data with laboratory results offered no additional gains (see Supplementary Table 11). For the unhealthy alcohol use screen task, where LLaVA-Med is not applicable, the published baseline performed better than the cross-attention and MoE methods, as expected given its training on a larger dataset. We also retrained a 1D-CNN model, following the approaches in the published baseline, on the same small cohort used by MoMA and observed that it underperformed (AUROC 0.641, AUPR 0.325, see Supplementary Table 10) relative to the published baseline model trained on the larger dataset. Despite these, MoMA achieved stronger results, with an AUROC of around 0.75 and an AUPR near 0.50.

To provide a comprehensive evaluation, we also reported macro-AUROC for the chest trauma severity and multitask chest and spine trauma stratification tasks, as well as F1 scores for the unhealthy alcohol use screening task in the Supplementary Note 1. These additional results are consistent with the trends shown in Figure 2.

Subgroup Analysis

Subgroup analysis is essential in clinical prediction models to ensure consistent performance across patient populations. By evaluating model performance across subpopulations, such as different racial and sex groups, researchers can identify variations in performance that may affect certain populations. The subgroup analysis results are presented in Figure 3 (see also Supplementary Table 4 and Supplementary Table 7). Because the limited number of non-white cases in the unhealthy alcohol use screening cohort led to high variance, we excluded the comparison for the white vs. non-white subgroups. Across all other subgroups, the MoMA architecture consistently achieved the best performance. Furthermore, we conducted paired t-tests to compare performance between females and males and between non-white and white individuals. MoMA demonstrated consistent performance across subgroups, while the baseline methods showed notable differences in subgroup performance in the chest trauma severity stratification and the multitask chest and spine trauma severity stratification tasks.

Ablation Study

To validate the contribution of non-text modalities to the improved performance, we conducted ablation studies in which non-text inputs were removed while keeping the remaining components of MoMA unchanged. Specifically, we excluded chest radiographs in the chest trauma severity stratification and the multitask chest and spine trauma severity stratification tasks and lab measurements in the unhealthy alcohol use screening cohort. The results are presented in Figure 4, with corresponding numerical values provided in Supplementary Table 8 for reference. MoMA with multimodal input outperformed its text-only counterparts. These results highlight that MoMA's improved performance is not only attributed to the enhanced text understanding capabilities of LLMs but also to the architecture's ability to effectively integrate and leverage non-text modalities.

Case Study

Figure 5 provides examples illustrating how the MoMA architecture effectively incorporated non-text information and synthesized all available data into an aggregated summary. In the upper example, the chest radiograph specialist agent analyzed radiographic images to identify or exclude severe conditions. In this case, the agent's output confirmed the absence of severe findings, enabling the architecture to accurately stratify the injury as moderate. However, the text-only approach incorrectly classified the case as severe, highlighting its limitation in utilizing critical radiographic information. In the lower example, the specialist agents and the aggregator agent collaboratively processed extensive clinical text (containing thousands of words) and lab measurements (comprising tens of measurements). These inputs were distilled into a concise, focused summary that preserved critical information while filtering out irrelevant details, which allowed the predictor agent to make an accurate classification while enhancing the transparency and interpretability of MoMA's decision-making process.

Discussion

In this work, we introduced Mixture-of-Multimodal-Agents (MoMA), a flexible architecture designed to harness the power of pretrained LLMs for clinical prediction tasks involving multimodal medical data. MoMA adopts a modular, plug-and-play design that enables state-of-the-art LLMs to serve as specialist agents for processing specific data types, allowing them to be easily swapped or extended based on task requirements. The proposed architecture was validated on three clinical tasks involving different combinations of EHR modalities (radiographs + clinical text and lab measurements + clinical text) and different prediction types (multiclass, multitask, and binary classification). Additionally, all the tasks utilized private datasets, ensuring that there was no risk of data leakage from the LLMs' pretraining phase, as publicly available datasets may have been incorporated into their pretraining phase. Across these tasks, MoMA achieved superior performance compared to baseline methods, showcasing its potential as a highly effective and flexible solution to handling a wide range of clinical prediction tasks.

Current studies on multi-agent architectures have largely emphasized their enhanced text understanding capabilities, particularly in generative tasks [60, 61, 62, 63]. However, they have yet to demonstrate whether such architectures can effectively leverage their text understanding abilities to improve prediction performance in classification tasks involving multimodal EHR data, which is a critical need in clinical applications. Our work directly addresses this gap by introducing MoMA, which harnesses the text understanding power of LLMs to achieve superior performance in clinical classification tasks across various modalities. To further illustrate its utility, we provided detailed case study results showcasing how MoMA's specialist agents extract and summarize key information from diverse input modalities. These agents contribute unique perspectives, which are then integrated by the aggregator agent to distill complex data into concise summaries. This sequential approach improves the transparency of the decision-making process. In clinical settings, where understanding the rationale behind predictions is as important as achieving high accuracy, MoMA offers an effective solution that combines performance with interpretability.

Unlike common multimodal integration methodologies that require extensive paired pretraining data, MoMA offers a significant advantage in its flexibility to utilize existing pretrained models for “projecting" non-text modalities into the text space. This independence allows MoMA to generalize better across diverse clinical tasks without being constrained by the availability of specific paired datasets during pretraining. By leveraging pretrained multimodal LLMs, MoMA encodes heterogeneous data types efficiently, combining their strengths in a unified architecture that is adaptable to varying input formats, such as sequential tabular data or variable numbers of medical images. Transforming all modalities to text also leverages the strengths of LLMs as language models that excel with textual input.

In this work, MoMA demonstrated superior performance in clinical tasks involving different combinations of EHR modalities: chest radiographs with clinical text, and lab measurements with clinical text. Leveraging recent advances in multimodal LLMs, MoMA can be readily adapted to various clinical scenarios by incorporating diverse non-text modalities beyond the chest radiographs and lab measurements presented here, simply by selecting appropriate specialist agents. For example, specialist agents such as BrainGPT[64] for 3D CT scans, ConcepPath[65] for histopathology images, and scGPT[66] for single-cell sequencing data can be easily integrated into the MoMA workflow when these modalities are involved. Moreover, MoMA's modular design facilitates the seamless integration of newer and more advanced LLMs as specialist, aggregator, or predictor agents, underscoring its flexibility and broad applicability.

In the chest trauma severity stratification and multitask chest and spine trauma severity stratification tasks, adding a third modality (e.g., lab measurement) to MoMA is feasible but did not result in performance improvement over the clinical text + radiographs setting. We attribute this to the already high performance achieved by the two modalities alone on these challenging multiclass tasks, suggesting that much of the relevant clinical signal may already be captured by these two rich modalities in this specific context. That said, MoMA supports the integration of different modalities with minimal overhead. This design align with our belief that clinical applications are inherently use-case specific: in some tasks, lab measurements may be critical; in others, they may be redundant or less informative.

Because our evaluation used private datasets, external researchers cannot directly access the original data. To address this and enhance reproducibility, we also demonstrated MoMA using publicly available datasets (MIMIC-IV[67] and MIMIC-CXR-JPG[68]), including admission notes and chest radiograph images, to predict in-hospital mortality in patients admitted to the Trauma Surgical Intensive Care Unit (TSICU). Instructions for reproducing MoMA's performance are available via the link provided in the "Code Availability" section.

Another key advantage of MoMA is that it doesn't require any training process for the non-text modality translation. By requiring fine-tuning only on the predictor agent, MoMA reduces the computational burden and data requirements typically associated with training large-scale multimodal models. This not only accelerates the training process but also makes MoMA more accessible for applications with limited resources, further demonstrating its versatility.

Hallucination [69] is a known limitation of LLMs. While directly addressing this issue is beyond the primary scope of our study and remains an open challenge in the field, we have taken steps to mitigate its potential impact within our classification pipeline. MoMA does not rely on the generated summaries for human interpretation. Instead, these summaries serve as intermediate representations used for downstream prediction. The final classifications are produced by feedforward layers applied to the state embeddings of the predictor agent, and the predictor agent is fine-tuned using groundtruth clinical labels. This training process encourages alignment between model predictions and true clinical outcomes, thereby reducing the negative impact from hallucinations that may arise in the generated summaries.

While MoMA shows promise for low resource with high accuracy predictions, several limitations need further exploration. First, the interactions between LLM agents in MoMA remain relatively simple, and enhancing communication and coordination of agents [70] could further improve the model's capabilities. Second, although MoMA offers flexibility in selecting pretrained LLMs in a plug-and-play manner, users should be aware that limitations such as introducing hallucination [69] and omitting necessary signals [71] are inherent to LLMs and may still impact classification performance, though MoMA is fine-tuned using groundtruth clinical labels. Finally, while this study focuses on clinical classification tasks, MoMA can be potentially extended to broader applications, such as medical visual question answering (Medical VQA) [72]; however, further validation is needed to support such extensions.

In summary, MoMA represents an advancement in leveraging the strength of LLMs with multimodal medical data for clinical prediction tasks. It demonstrates superior performance compared to state-of-the-art methods, meanwhile offering interpretability, computational efficiency, and flexibility to diverse input formats, making it a highly promising tool for improving clinical decision making.

Methods

In this section, we present our proposed MoMA architecture. We begin with outlining the foundational design insights behind MoMA. Then we describe the roles of each agent and their collaborative processes in generating the final output in detail. In addition, we introduce the Cross-attention and Mixture-of-Experts based fusion methods which serve as comparison baselines. We adhere to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - Large Language Models (TRIPOD-LLM) guidelines [73] and completed the accompanying checklist, meeting all reporting requirements, as detailed in Supplementary Note 4.

Mixture-of-Multimodal-Agents Architecture

The design of the MoMA architecture draws upon two frameworks: Mixture-of-Experts (MoE) and Mixture-of-Agents (MoA). The Mixture-of-Experts (MoE) framework [74] is a well-established approach for enhancing model performance by selectively activating specialized "experts". The basic mechanism behind MoE involves a gating network that learns to route each input to the most relevant experts, thus enabling the model to adaptively leverage specialist knowledge without overloading each expert with irrelevant information. The MoE approach has demonstrated excellent performance across a range of tasks within natural language processing [75] and computer vision domains.

Building on this foundation, the Mixture-of-Agents (MoA) architecture [38] extends the MoE concept by replacing traditional experts and the gating network with LLM agents, thereby leveraging the distinct expertise of different pretrained LLMs. It also harnesses the collaborativeness of LLMs, wherein an LLM agent typically achieves better performance when it incorporates the output of another LLM agent as additional input, compared to relying solely on the original input text. The MoA framework achieves state-of-the-art performance in text understanding and summarization tasks.

In this study, we extend the MoA framework to accommodate multiple modalities in EHR data, proposing the MoMA architecture. Unlike the original MoA which employs multiple LLM agents to analyze a single text input, MoMA assigns a dedicated specialist agent to convert each non-text modality into a text summary. These summaries from various modalities are then processed through a stack of LLMs to leverage their collaborative capabilities. This architecture can support a growing number of modalities through the integration of multimodal LLMs in a plug-and-play manner. Building upon this foundational insight, we now describe the specific details of the MoMA architecture.

Figure 1 illustrates the architecture of MoMA. As one of the core design principles of MoMA is to align various modalities to the text space, the original clinical text remains unprocessed until it reaches the aggregator agent. Specialist agents convert each non-text modality (e.g., images, structured lab results) into concise textual summaries. For example, multimodal LLMs such as LLaVA-Med [23] and CXR-LLAVA [34] can generate text summaries from medical images, while structured EHR data can be summarized using general-purpose LLMs [76] like Llama-3 [77]. The resulting text summaries from non-text modalities are concatenated with the original clinical notes, and this concatenated text is then provided as input to the aggregator agent. Formally, for sample i, the input to the aggregator agent is

$$m_i = t_i \oplus \left(\bigoplus_{j=1}^{M} \mathcal{S}_j(k_{i,j}) \right)$$

where t_i is the original clinical text of sample i, $\oplus$ denotes text concatenation, M denotes number of non-text modalities, $\bigoplus_{j=1}^{M}$ represents the operation of text concatenation across the summaries generated from the M non-text modalities, $\mathcal{S}_j$ is the specialist agent handling the jth non-text modality, $k_{i,j}$ is the input data for the jth non-text modality of sample i.

The aggregator agent receives m_i as input and is prompted to generate a comprehensive and concise summary that integrates information from all available modalities. This summary is then passed to the predictor agent, which uses it to make predictions. Formally, these steps are represented as follows:

$$\hat{y}_i = \mathcal{P}(\mathcal{A}(m_i))$$

where $\mathcal{A}$ is the aggregator agent, $\mathcal{P}$ is the predictor agent, and $\hat{y}_i$ is the output prediction of sample i. In our study, we utilize Llama-3 as the specialist agent for both free text and tabular data, while CXR-LLAVA serves as the specialist agent for chest radiographs. Additionally, Llama-3 is used as both the aggregator and predictor agent. To generate the final prediction, we extract the hidden state corresponding to the last token output by the aggregator agent (fine-tuned Llama-3) and pass it through a feedforward layer to produce the final logit.

Prompt Engineering for MoMA

We provide general guidelines for creating prompts tailored to the MoMA framework. Users can input the elements listed below into an LLM to automatically generate these prompts. Specifically, prompts should guide the text specialist agent to extract task-specific clinical information, instruct the non-text specialist agent to identify data that is both relevant to and complementary to clinical text, and direct the aggregator agent to synthesize and summarize outputs from all specialist agents. Example prompts for chest trauma severity stratification and unhealthy alcohol use screening are provided in Supplementary Note 3.

Outlined below are the prompting guidelines for the text specialist agent.

1. **Identify Relevant Points**
   - Read the entire note carefully. Highlight any direct mentions related to your primary focus.
   - *Example: "As an experienced trauma physician...summarizing chest trauma injuries."*
2. **Apply Specified Criteria**
   - Compare the details you've noted to any stated thresholds (e.g., a certain number of drinks per day, specific radiology report findings).
   - *Example: "Excessive Consumption: four or more drinks in a single day if female, five or more if male..."*
3. **Incorporate Additional Patient Attributes**
   - Look for coexisting conditions or factors that might explain or overlap with your primary findings (e.g., other diagnoses, lifestyle issues).
   - *Example: "Scan the note for conditions...like hepatitis, renal dysfunction..."*

4. **Maintain Clarity and Flow**
    - Present the findings in one coherent summary. Start with definitive evidence, then mention borderline or uncertain items, and end with confounders.
    - *Example: "Summarize your findings by clearly separating evidence of [X] from evidence of other conditions..."*
5. **Professional Language and Confidentiality**
    - Use objective, clinical terminology
    - Omit Protected Health Information (PHI) and avoid speculation.
    - *Example: "Do not include any Protected Health Information (PHI)... Do not guess or infer based on information that does not directly prove it."*
6. **Final Structured Review**
    - Conclude with a succinct overview of confirmed evidence, partial/absent findings, and relevant confounders.
    - *Example:"If no direct evidence meeting these criteria is found, explicitly state that no direct evidence is found."*

The following outlines the prompting guidelines for the non-text specialist agent.

1. **Identify Relevant Indicators**
    - Scan for direct measures relevant to the outcome (e.g., blood alcohol levels).
    - Look for key values linked to the condition or focus of your review.
    - *Example: "Identify any initial measurements commonly linked to alcohol consumption..."*
2. **Evaluate Indirect Evidence**
    - Check secondary or indirect measures relevant to the outcome (e.g., abnormal enzyme levels).
    - Note whether these indirectly suggest the issue at hand.
    - *Example: "Consider labs with indirect evidence... elevated liver enzymes..."*
3. **Summarize Concisely**
    - Create a short, clear overview highlighting the most pertinent findings.
    - Maintain professionalism and exclude any unnecessary speculation.

The guidelines for prompting the aggregator specialist agent are provided below.

1. **Gather Key Points from Agent-Generated Summaries**
    - Collect all relevant findings from each summary, such as clinical observations, lab findings, or imaging results.
    - *Example: "Review the agent-generated summaries to identify any details related to alcohol use, including behavior patterns..."*
2. **Handle Contradicted Information**
    - If the reports are contradictory, do not let automatically generated information overwrite established confirmed evidence.
    - *Example: "Ensure that the LLM-generated radiology reports do not override clinical notes in cases where they contradict each other..."*
3. **Exclude Confounding Information**
    - Exclude any details if they have an alternative cause not relevant to the goal.
    - *Example: "For any mentioned lab abnormalities, review the clinical summaries to determine if they may have causes unrelated to alcohol use."*
4. **Create a Unified Summary**
    - Generate a comprehensive summary by integrating agent-generated reports from multimodal data for specific prediction tasks.
    - *Example: "Write a comprehensive summary of the patient's alcohol use, integrating relevant details from both the clinical summaries and lab results."*

LLaVA-Med

We adopt the LLaVA-Med v1.5 (Mistral-7B) checkpoint trained in April 2024 as a medical vision-language large model benchmark [23]. This model extends the general-purpose LLaVA framework to the medical domain by adapting it for radiograph interpretation and multimodal instruction following.

LLaVA-Med integrates a CLIP image encoder [11] with a Mistral-7B language model [78] via a learned projection layer that aligns visual embeddings to the language model's input space. This model extends the general-purpose LLaVA framework [79] to the medical domain by adapting it for radiograph interpretation and multimodal instruction following.

Following the classification setup used in the original LLaVA-Med paper, we treat the prediction task as a generative process. Specifically, we prompt the model with a clinical query and ask it to choose from a predefined list of candidate class labels (e.g., negative, moderate, severe). The final prediction is determined by string matching of the model's response.

To address the known performance degradation with long or noisy textual input for medical vision-language large models [80, 81], we followed the strategy of Thawkar et al.[82], using a separate large language model to first summarize the full clinical notes into concise descriptions of trauma-related information. The resulting summary is then paired with the chest X-ray image and provided to LLaVA-Med for fine-tuning.

We fine-tuned LLaVA-Med on the development set to enhance predictive performance. During training, we supervised the model using paired inputs (chest radiographs + summarized clinical notes) and corresponding class labels as instruction–response pairs. The prompts used for this benchmark have been presented in Supplementary Note 3.

Cross-attention Based Fusion

Cross-attention provides a robust framework for fusing encoded representations from two modalities by allowing one modality to selectively attend to relevant features in the other. Let U_a and U_b denote the encoded representations of two modalities, where U_a provides the primary signals for classification, and U_b serves as the complementary modality. The cross-attention mechanism computes a refined representation of U_a by attending to U_b as follows.

It first conducts linear transformations that project U_a and U_b into query (Q), key (K), and value (V) spaces:

$$\begin{aligned} Q &= U_a W_Q, \\ K &= U_b W_K, \\ V &= U_b W_V \end{aligned}$$

where W_Q, W_K, W_V are learnable weight matrices.

Then scaled dot-product attention computes the alignment scores between Q and K:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right),$$

where A is the attention weight matrix, normalized row-wise to ensure row-wise weights sum to 1.

The values V are aggregated based on the attention weights to produce the fused representation F_a:

$$F_a = AV,$$

where F_a is the updated representation of U_a, enriched with relevant information from U_b. The refined representation F_a is combined with the original U_a via residual connection to form the final fused representation for downstream tasks.

$$F = F_a + U_a.$$

Sparse Mixture-of-Experts (MoE) Based Fusion

The Sparse Mixture-of-Experts (MoE) model is an efficient approach for fusing encoded representations from M different EHR modalities, denoted by $\{U_1, U_2, \ldots, U_M\}$, including both text and non-text modalities. The process is outlined as follows:

The MoE framework includes a set of K expert networks $\{E_1, E_2, \ldots, E_K\}$. Each expert specializes in processing subsets of the input space, enabling the model to handle diverse patterns across modalities. Each expert processes one or more encoded representations U_m based on the routing mechanism. A gating network determines how to route the input representations U_m to the experts. The gating network generates scores $G_{m,k}$ for each combination of modality m and expert k:

$$G_{m,k} = \text{softmax}(\phi(U_m)),$$

where ϕ is a neural network applied to U_m.

Then the outputs from the active experts are aggregated to form the final fused representation F.

$$F = \sum_{k \in \text{active}} G_{m,k} \cdot E_k(U_m).$$

The fused representation F is passed to a final prediction layer to perform the downstream classification task.

Published SOTA

Following the design proposed by Gao et al.[5], we fine-tuned ClinicalBERT using clinical text inputs, specifically ED notes and radiology reports. Due to ClinicalBERT's input length limit, notes that are more likely to provide comprehensive information were prioritized: ED notes were sorted from longest to shortest, and radiology reports from earliest to latest. The first 300 tokens were allocated to ED notes, with the remaining tokens assigned to radiology reports. If one note type used fewer tokens than its allocation, the unused tokens were reallocated to the other. Notes exceeding the overall token limit were truncated. This ClinicalBERT baseline was finetuned on the development set of this work and validated on the test set independently. We used this baseline for the chest trauma severity and the multitask chest and spine trauma severity stratification tasks.

In the work by Afshar et al.[50], clinical concepts, such as diseases, symptoms, anatomical sites, medications, and procedures, were extracted from all available notes within each patient encounter using the NLP engine, clinical Text And Knowledge Extraction System (cTAKES)[83]. These extracted concepts were then embedded and processed by a one-dimensional convolutional neural network (1D-CNN) for multi-task predictions of alcohol, opioid, and non-opioid misuse. We utilized the trained model from [50] and reported the performance of its unhealthy alcohol use prediction head on our test set without any additional fine-tuning. In addition, we retrained the model on the same cohort used by MoMA for the unhealthy alcohol use screening task.

Computational Resources and Runtime

All experiments were conducted using two A100 GPUs with 80 GB of memory. For the chest trauma severity stratification task and the multitask chest and spine trauma severity stratification task, generating summaries with specialist agents takes approximately 72 hours, while the remaining processes are completed in under 4 hours. For the unhealthy alcohol use screening task, specialist summary generation took around 48 hours, with the remaining processes completed in under 3 hours.

Ethics Statement

This study was approved by the University of Wisconsin-Madison Minimal Risk Research Institutional Review Board (IRB) under the following protocols. For the chest trauma severity stratification and multitask chest and spine trauma severity stratification tasks, ethical approval was granted under IRB protocol #2019-1258 with a waiver of informed consent. For the unhealthy alcohol use screening task, ethical approval was granted under IRB protocol #2021-0509. Informed consent was obtained from all participants.

Data Availability

Due to legal and regulatory constraints, the data utilized in this study are not publicly accessible. Our data was obtained from the UW Health system after receiving approval from the IRB. Our data use agreements do not permit sharing clinical data. Researchers with an interest in accessing the data can reach out to the corresponding authors or Madeline Oguss at mkoguss@medicine.wisc.edu. A demonstration of MoMA, using the publicly available MIMIC-IV and MIMIC-CXR datasets, is provided to ensure the reproducibility of our methods.

The code for this project, including the chest trauma severity stratification and unhealthy alcohol use screening tasks as well as the demonstration using the MIMIC-IV and MIMIC-CXR datasets, is available at `https://git.doit.wisc.edu/smph-public/dom/uw-icu-data-science-lab-public/moma`

Acknowledgments

This was supported in part by the National Institute of Health/National Institute on Drug Abuse R01DA051464 (MA, MO), National Library of Medicine R01LM012973 (MA) and the National Science Foundation DMS-2054346 (JG, GC).

Author Contributions

G. C. and M. A. conceived the study; J. G. developed the methods with input from G. C. and M.A.; J. G. conducted the simulations and application with help from J. C.; A.R., R. B., A. S., A. M. assisted result interpretation; G. C., M.A. and M. C. supervised the study; G. C. and M.A. provided funding, J. G., G. C. and M. A. wrote the draft. All authors reviewed and provided revision input on the manuscript.

Conflicts of Interest

The authors have no competing interests to declare.

References

[1] Qiong Cai et al. "A survey on multimodal data-driven smart healthcare systems: approaches and applications". In: *IEEE Access* 7 (2019), pp. 133583–133599.

[2] Benjamin Rohaut et al. "Multimodal assessment improves neuroprognosis performance in clinically unresponsive critical-care patients with brain injury". In: *Nature Medicine* (2024), pp. 1–7.

[3] Luis R Soenksen et al. "Integrated multimodal artificial intelligence framework for healthcare applications". In: *NPJ digital medicine* 5.1 (2022), p. 149.

[4] Caleb Winston et al. "Multimodal Clinical Prediction with Unified Prompts and Pretrained Large-Language Models". In: *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*. IEEE. 2024, pp. 679–683.

[5] Jifan Gao et al. "Automated stratification of trauma injury severity across multiple body regions using multi-modal, multi-class machine learning models". In: *Journal of the American Medical Informatics Association* 31.6 (2024), pp. 1291–1302.

[6] Adrienne Kline et al. "Multimodal machine learning in precision health: A scoping review". In: *npj Digital Medicine* 5.1 (2022), p. 171.

[7] Julián N Acosta et al. "Multimodal biomedical AI". In: *Nature Medicine* 28.9 (2022), pp. 1773–1784.

[8] Shih-Cheng Huang et al. "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines". In: *NPJ digital medicine* 3.1 (2020), p. 136.

[9] Junnan Li et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation". In: *International conference on machine learning*. PMLR. 2022, pp. 12888–12900.

[10] Xing Han et al. "Fusemoe: Mixture-of-experts transformers for fleximodal fusion". In: *arXiv preprint arXiv:2402.03226* (2024).

[11] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[12] Shruthi Bannur et al. "Learning to exploit temporal structure for biomedical vision-language processing". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15016–15027.

[13] Noah Cohen Kalafut, Xiang Huang, and Daifeng Wang. "Joint variational autoencoders for multimodal imputation and embedding". In: *Nature Machine Intelligence* 5.6 (2023), pp. 631–642.

[14] Valerio Guarrasi et al. "A Systematic Review of Intermediate Fusion in Multimodal Deep Learning for Biomedical Applications". In: *arXiv preprint arXiv:2408.02686* (2024).

[15] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. "Multimodal deep learning for biomedical data fusion: a review". In: *Briefings in Bioinformatics* 23.2 (2022), bbab569.

[16] Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. "MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images". In: *Machine Learning for Healthcare Conference*. PMLR. 2022, pp. 479–503.

[17] Junnan Li et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models". In: *International conference on machine learning*. PMLR. 2023, pp. 19730–19742.

[18] Jean-Baptiste Alayrac et al. “Flamingo: a visual language model for few-shot learning”. In: *Advances in neural information processing systems* 35 (2022), pp. 23716–23736.

[19] Zhiliang Peng et al. “Kosmos-2: Grounding multimodal large language models to the world”. In: *arXiv preprint arXiv:2306.14824* (2023).

[20] Danny Driess et al. “Palm-e: An embodied multimodal language model”. In: (2023).

[21] Rohit Girdhar et al. “Imagebind: One embedding space to bind them all”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 15180–15190.

[22] Jiaming Han et al. “Onellm: One framework to align all modalities with language”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 26584–26595.

[23] Chunyuan Li et al. “Llava-med: Training a large language-and-vision assistant for biomedicine in one day”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[24] Vishwesh Nath et al. “Vila-m3: Enhancing vision-language models with medical expert knowledge”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 14788–14798.

[25] Sunan He et al. “GSCo: Towards Generalizable AI in Medicine via Generalist-Specialist Collaboration”. In: *arXiv preprint arXiv:2404.15127* (2024).

[26] Kevin M Boehm et al. “Harnessing multimodal data integration to advance precision oncology”. In: *Nature Reviews Cancer* 22.2 (2022), pp. 114–126.

[27] Jee Suk Chang et al. “Continuous multimodal data supply chain and expandable clinical decision support for oncology”. In: *npj Digital Medicine* 8.1 (2025), p. 128.

[28] Wayne Xin Zhao et al. “A survey of large language models”. In: *arXiv preprint arXiv:2303.18223* (2023).

[29] Jason Wei et al. “Emergent abilities of large language models”. In: *arXiv preprint arXiv:2206.07682* (2022).

[30] Arun James Thirunavukarasu et al. “Large language models in medicine”. In: *Nature medicine* 29.8 (2023), pp. 1930–1940.

[31] Hanyin Wang et al. “DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients”. In: *npj Digital Medicine* 7.1 (2024), p. 16.

[32] Fenglin Liu et al. “A medical multimodal large language model for future pandemics”. In: *NPJ Digital Medicine* 6.1 (2023), p. 226.

[33] Bowen Gu et al. “Probabilistic Medical Predictions of Large Language Models”. In: *arXiv preprint arXiv:2408.11316* (2024).

[34] Seowoo Lee et al. “Cxr-llava: Multimodal large language model for interpreting chest x-ray images”. In: *arXiv preprint arXiv:2310.18341* (2023).

[35] Yinghao Zhu et al. “Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data”. In: *arXiv preprint arXiv:2402.01713* (2024).

[36] Yanjun Gao et al. “When Raw Data Prevails: Are Large Language Model Embeddings Effective in Numerical Data Representation for Medical Machine Learning Applications?” In: *arXiv preprint arXiv:2408.11854* (2024).

[37] Gary Lupyan. “The centrality of language in human cognition”. In: *Language Learning* 66.3 (2016), pp. 516–553.

[38] Junlin Wang et al. “Mixture-of-Agents Enhances Large Language Model Capabilities”. In: *arXiv preprint arXiv:2406.04692* (2024).

[39] Hanzhou Li et al. “Ethics of large language models in medicine and medical research”. In: *The Lancet Digital Health* 5.6 (2023), e333–e335.

[40] Joschka Haltaufderheide and Robert Ranisch. “The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs)”. In: *NPJ digital medicine* 7.1 (2024), p. 183.

[41] Juan P Herrera-Escobar and Jeffrey C Schneider. “From survival to survivorship—framing traumatic injury as a chronic condition”. In: *The New England journal of medicine* 387.7 (2022), p. 581.

[42] R Lefering et al. “Epidemiology of in-hospital trauma deaths”. In: *European journal of trauma and emergency surgery* 38 (2012), pp. 3–9.

[43] Anna Granström et al. “A criteria-directed protocol for in-hospital triage of trauma patients”. In: *European Journal of Emergency Medicine* 25.1 (2018), pp. 25–31.

[44] Shahid Shafi et al. “The trauma quality improvement program of the American College of Surgeons Committee on Trauma”. In: *Journal of the American College of Surgeons* 209.4 (2009), 521–530e1.

[45] Cameron S Palmer, Belinda J Gabbe, and Peter A Cameron. “Defining major trauma using the 2008 Abbreviated Injury Scale”. In: *Injury* 47.1 (2016), pp. 109–115.

[46] Wu Jinyi et al. “Global, regional, and national mortality of tuberculosis attributable to alcohol and tobacco from 1990 to 2019: A modelling study based on the Global Burden of Disease study 2019”. In: *Journal of Global Health* 14 (2024).

[47] Simon Coulton. “Alcohol misuse”. In: *BMJ Clinical Evidence* 2011 (2011).

[48] Jennifer McNeely et al. “Performance of the tobacco, alcohol, prescription medication, and other substance use (TAPS) tool for substance use screening in primary care patients”. In: *Annals of internal medicine* 165.10 (2016), pp. 690–699.

[49] Emily Alsentzer et al. “Publicly available clinical BERT embeddings”. In: *arXiv preprint arXiv:1904.03323* (2019).

[50] Majid Afshar et al. “Development and multimodal validation of a substance misuse algorithm for referral to treatment using artificial intelligence (SMART-AI): a retrospective deep learning study”. In: *The Lancet Digital Health* 4.6 (2022), e426–e435.

[51] Danfeng Guo and Demetri Terzopoulos. “Prompting Medical Large Vision-Language Models to Diagnose Pathologies by Visual Question Answering”. In: *arXiv preprint arXiv:2407.21368* (2024).

[52] Kangyu Zhu et al. “Guiding Medical Vision-Language Models with Explicit Visual Prompts: Framework Design and Comprehensive Exploration of Prompt Variations”. In: *arXiv preprint arXiv:2501.02385* (2025).

[53] Xikai Yang et al. “Medical Large Vision Language Models with Multi-Image Visual Ability”. In: *arXiv preprint arXiv:2505.19031* (2025).

[54] Lihua Jian et al. “Rethinking Cross-Attention for Infrared and Visible Image Fusion”. In: *arXiv preprint arXiv:2401.11675* (2024).

[55] Wenhao Zheng et al. “Multimodal clinical trial outcome prediction with large language models”. In: *arXiv preprint arXiv:2402.06512* (2024).

[56] Yanjun Gao et al. *Clinical natural language processing for secondary uses*. 2024.

[57] Wei Ouyang et al. “Analysis of the human protein atlas image classification competition”. In: *Nature methods* 16.12 (2019), pp. 1254–1261.

[58] Timothy Bergquist et al. “Evaluation of crowdsourced mortality prediction models as a framework for assessing artificial intelligence in medicine”. In: *Journal of the American Medical Informatics Association* 31.1 (2023), pp. 35–44.

[59] Timothy Bergquist et al. “A framework for future national pediatric pandemic respiratory disease severity triage: the HHS pediatric COVID-19 data challenge”. In: *Journal of Clinical and Translational Science* 7.1 (2023), e175.

[60] Bowen Jiang et al. “Multi-modal and multi-agent systems meet rationality: A survey”. In: *ICML 2024 Workshop on LLMs and Cognition*. 2024.

[61] Yilun Du et al. “Improving factuality and reasoning in language models through multiagent debate”. In: *arXiv preprint arXiv:2305.14325* (2023).

[62] Xiangru Tang et al. “Medagents: Large language models as collaborators for zero-shot medical reasoning”. In: *arXiv preprint arXiv:2311.10537* (2023).

[63] Junyou Li et al. “More agents is all you need”. In: *arXiv preprint arXiv:2402.05120* (2024).

[64] Cheng-Yi Li et al. “Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation”. In: *Nature Communications* 16.1 (2025), p. 2258.

[65] Weiqin Zhao et al. “Aligning knowledge concepts to whole slide images for precise histopathology image analysis”. In: *npj Digital Medicine* 7.1 (2024), p. 383.

[66] Haotian Cui et al. “scGPT: toward building a foundation model for single-cell multi-omics using generative AI”. In: *Nature Methods* 21.8 (2024), pp. 1470–1480.

[67] Alistair EW Johnson et al. “MIMIC-IV, a freely accessible electronic health record dataset”. In: *Scientific data* 10.1 (2023), p. 1.

[68] Alistair EW Johnson et al. “MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs”. In: *arXiv preprint arXiv:1901.07042* (2019).

[69] Lei Huang et al. “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions”. In: *ACM Transactions on Information Systems* 43.2 (2025), pp. 1–55.

[70] Adam Fourney et al. “Magentic-one: A generalist multi-agent system for solving complex tasks”. In: *arXiv preprint arXiv:2411.04468* (2024).

[71] Davide Caffagni et al. “The revolution of multimodal large language models: a survey”. In: *arXiv preprint arXiv:2402.12451* (2024).

[72] Asma Ben Abacha et al. “Vqa-med: Overview of the medical visual question answering task at imageclef 2019”. In: *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019. 2019.

[73] Jack Gallifant et al. “The TRIPOD-LLM reporting guideline for studies using large language models”. In: *Nature Medicine* (2025), pp. 1–10.

[74] Noam Shazeer et al. “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer”. In: *arXiv preprint arXiv:1701.06538* (2017).

[75] Niklas Muennighoff et al. “MTEB: Massive Text Embedding Benchmark”. In: *arXiv preprint arXiv:2210.07316* (2022). DOI: 10.48550/ARXIV.2210.07316. URL: https://arxiv.org/abs/2210.07316.

[76] Wenqi Shi et al. “Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records”. In: *ICLR 2024 Workshop on Large Language Model (LLM) Agents*. 2024.

[77] Abhimanyu Dubey et al. “The llama 3 herd of models”. In: *arXiv preprint arXiv:2407.21783* (2024).

[78] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

[79] Haotian Liu et al. “Visual instruction tuning”. In: *Advances in neural information processing systems* 36 (2023), pp. 34892–34916.

[80] Peng Xia et al. “Mmed-rag: Versatile multimodal rag system for medical vision language models”. In: *arXiv preprint arXiv:2410.13085* (2024).

[81] Peng Xia et al. “Rule: Reliable multimodal rag for factuality in medical vision language models”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024, pp. 1081–1093.

[82] Omkar Thawkar et al. “Xraygpt: Chest radiographs summarization using medical vision-language models”. In: *arXiv preprint arXiv:2306.07971* (2023).

[83] Guergana K Savova et al. “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications”. In: *Journal of the American Medical Informatics Association* 17.5 (2010), pp. 507–513.

Table 1: Cohort characteristics and label distributions. The columns under *“Trauma severity stratification"* present statistics of both the chest trauma severity stratification and the multitask chest and spine trauma severity stratification task, as these two tasks were conducted on the same cohort.

	Trauma severity stratification		Unhealthy alcohol use screening	
	Development	Test	Development	Test
Encounters, n	3,451	870	1,576	482
Age, median (IQR)	58 (37,74)	62 (42,77)	63 (50,73)	64 (51,73)
Total hours of stay, median (IQR)	104 (59,189)	104 (52,190)	100 (65,175)	144 (84,253)
Female, n (%)	1,303 (37.8)	346 (39.8)	744 (47.2)	207 (42.9)
White, n (%)	3,180 (92.1)	786 (90.3)	1,418 (90.0)	432 (89.6)
Labels for chest trauma severity stratification				
Negative	2,184 (63.3)	538 (61.8)	-	-
Minor/moderate	398 (11.5)	97 (11.1)	-	-
Serious or greater	869 (25.1)	235 (27.0)	-	-
Labels for spine trauma severity stratification				
Negative	2,394 (69.4)	626 (72.0)	-	-
Minor/moderate	773 (22.4)	187 (21.5)	-	-
Serious or greater	284 (8.2)	57 (6.6)	-	-
Labels for unhealthy alcohol use screening				
Negative	-	-	235 (14.9)	78 (16.2)
Positive	-	-	1,341 (85.1)	404 (83.8)

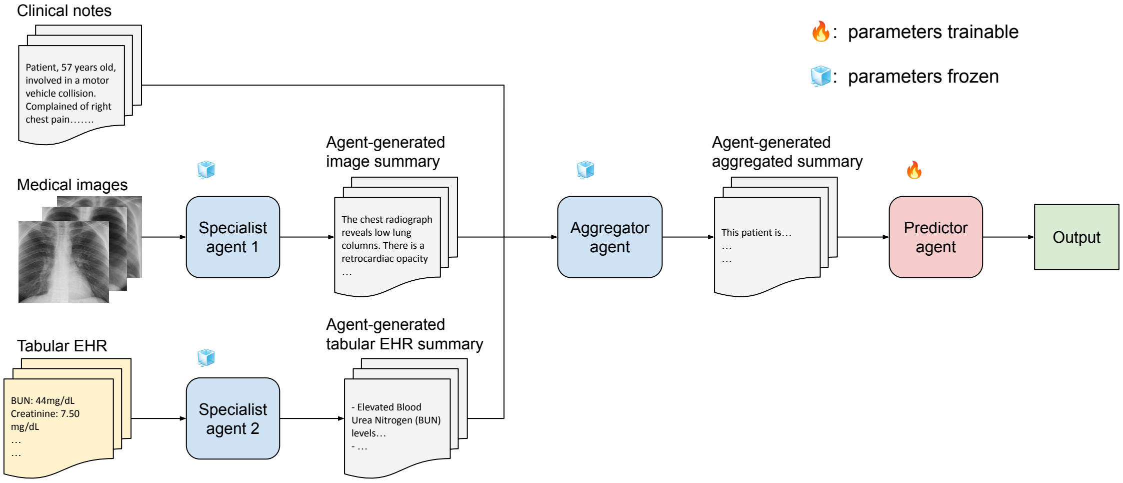

Clinical notes
Patient, 57 years old, involved in a motor vehicle collision. Complained of right chest pain.......
Medical images
Tabular EHR
BUN: 44mg/dL
Creatinine: 7.50 mg/dL
...
...
Specialist agent 1
Specialist agent 2
Agent-generated image summary
The chest radiograph reveals low lung columns. There is a retrocardiac opacity
...
Agent-generated tabular EHR summary
- Elevated Blood Urea Nitrogen (BUN) levels...
- ...
Aggregator agent
Agent-generated aggregated summary
This patient is...
...
...
Predictor agent
Output
🔥: parameters trainable
🧊: parameters frozen

Figure 1: Architecture of MoMA. Non-plain-text modalities, including medical images and tabular EHR data, are transformed into text summaries by specialist agents. These agent-generated summaries are then appended to the original clinical notes and passed to the aggregator agent, which produces a comprehensive and concise summary from the combined text. This aggregated summary is subsequently fed into a predictor agent to generate output predictions. Given the well-known capabilities of LLMs in clinical summarization, both the specialist and aggregator agents operate in a zero-shot setting, with only the predictor agent requiring training or fine-tuning, which reduces computational costs during the training phase.

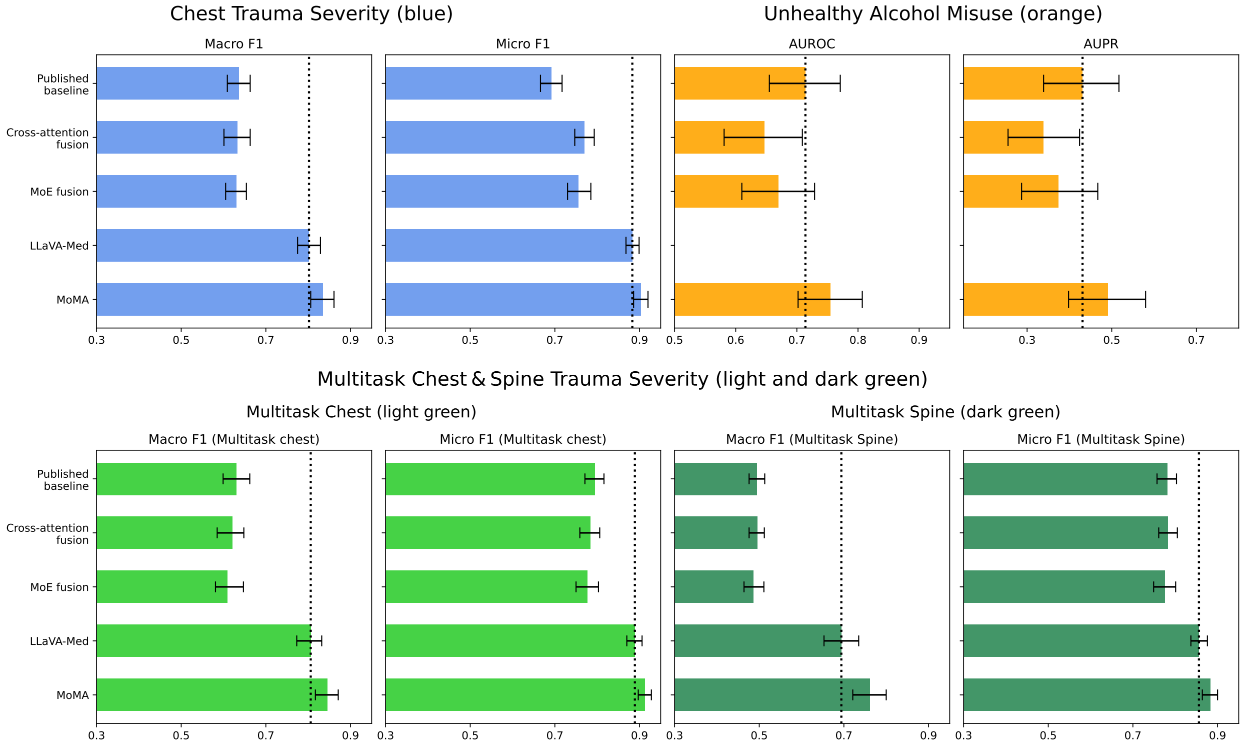


Figure 2: Comparison of discriminative performance and 95% confidence intervals between MoMA, the published SOTA baseline, and the cross-attention, MoE-based fusion, and fine-tuned LLaVA-Med. Dotted lines mark the strongest baselines, LLaVA-Med for chest trauma and multitask chest and spine trauma stratification, and the published baseline for the unhealthy alcohol use screening. Confidence intervals were calculated using bootstrap methods. Note that LLaVA-Med accepts only text + image inputs, so it cannot be evaluated on the unhealthy alcohol use screening task, which combines text with tabular data. MoMA consistently outperforms all applicable baselines across these clinical tasks.

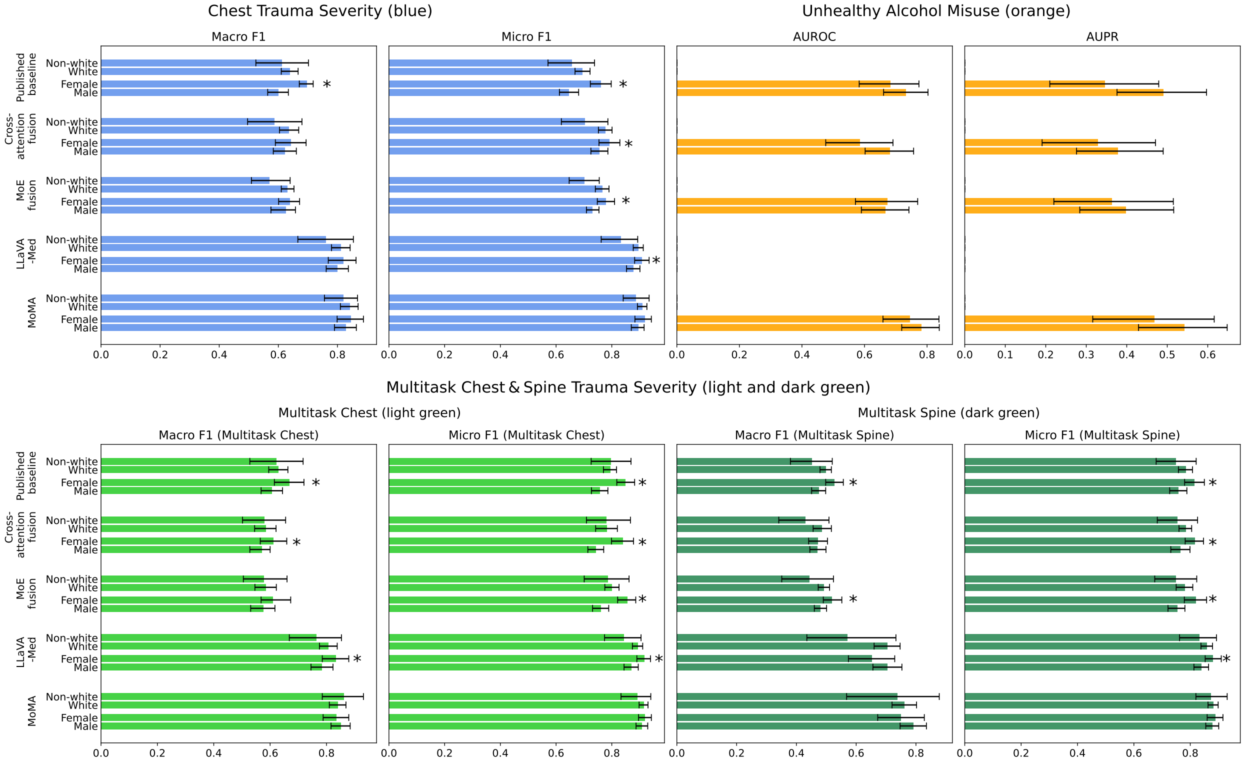


Figure 3: Comparison of model performance across racial and sex subgroups. We excluded the comparison for the white vs. non-white subgroups in the unhealthy alcohol use screening task due to the limited number of cases in the non-white subgroup. Statistically significant differences in performance within each group are marked with asterisks. MoMA achieves the highest and consistent performance across all subgroups in these clinical tasks.

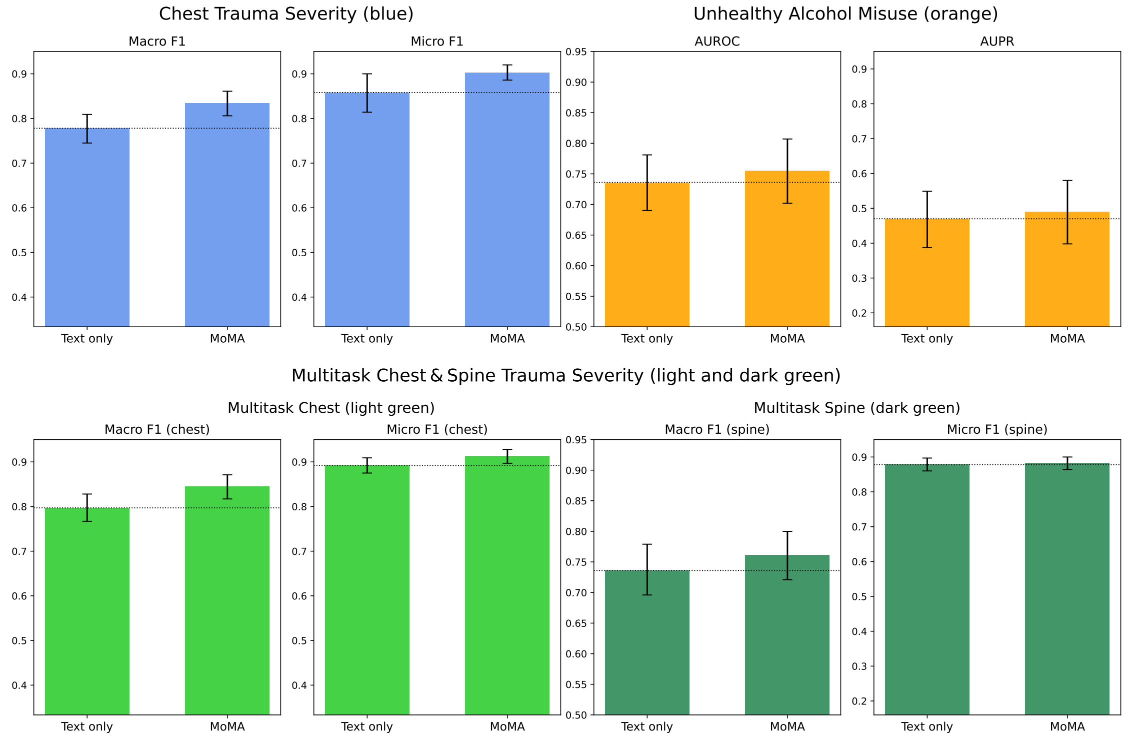
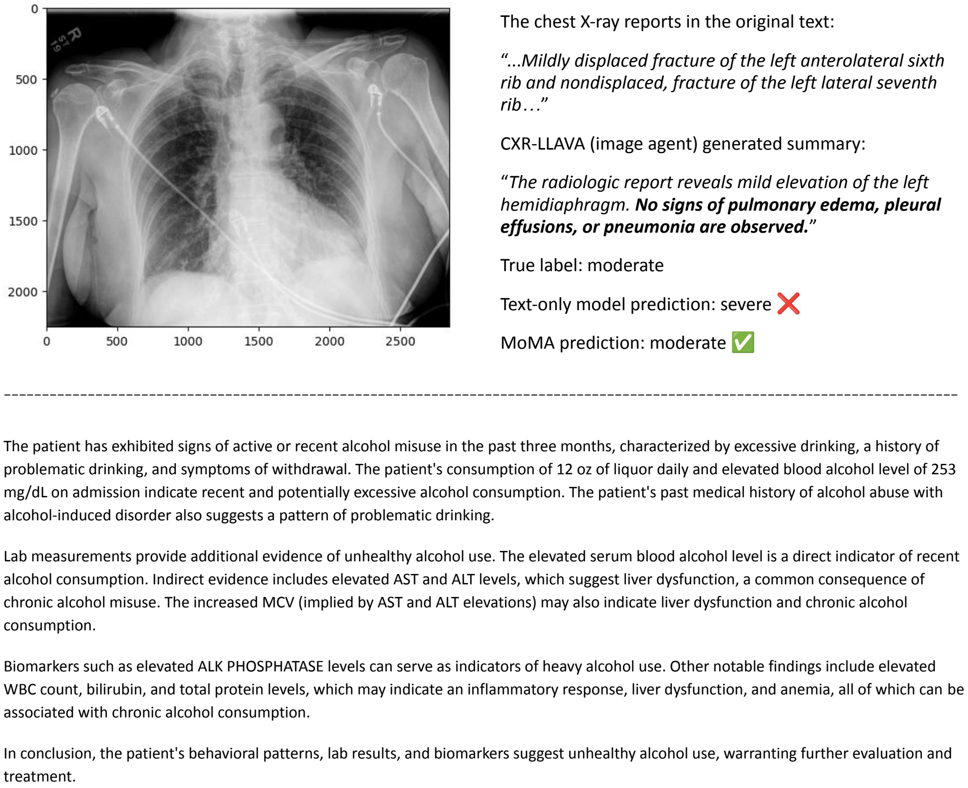


Figure 4: Ablation study evaluating the impact of removing non-text modalities from MoMA's input. MoMA consistently achieves superior performance compared to its text-only counterparts, demonstrating that performance improvements arise from the effective utilization of non-text modalities.



The chest X-ray reports in the original text:

"*...Mildly displaced fracture of the left anterolateral sixth rib and nondisplaced, fracture of the left lateral seventh rib…*"

CXR-LLAVA (image agent) generated summary:

"*The radiologic report reveals mild elevation of the left hemidiaphragm.* ***No signs of pulmonary edema, pleural effusions, or pneumonia are observed.***"

True label: moderate

Text-only model prediction: severe ❌

MoMA prediction: moderate ✅

The patient has exhibited signs of active or recent alcohol misuse in the past three months, characterized by excessive drinking, a history of problematic drinking, and symptoms of withdrawal. The patient's consumption of 12 oz of liquor daily and elevated blood alcohol level of 253 mg/dL on admission indicate recent and potentially excessive alcohol consumption. The patient's past medical history of alcohol abuse with alcohol-induced disorder also suggests a pattern of problematic drinking.

Lab measurements provide additional evidence of unhealthy alcohol use. The elevated serum blood alcohol level is a direct indicator of recent alcohol consumption. Indirect evidence includes elevated AST and ALT levels, which suggest liver dysfunction, a common consequence of chronic alcohol misuse. The increased MCV (implied by AST and ALT elevations) may also indicate liver dysfunction and chronic alcohol consumption.

Biomarkers such as elevated ALK PHOSPHATASE levels can serve as indicators of heavy alcohol use. Other notable findings include elevated WBC count, bilirubin, and total protein levels, which may indicate an inflammatory response, liver dysfunction, and anemia, all of which can be associated with chronic alcohol consumption.

In conclusion, the patient's behavioral patterns, lab results, and biomarkers suggest unhealthy alcohol use, warranting further evaluation and treatment.

Figure 5: The upper section showcases the output of the chest X-ray specialist agent, which excludes the presence of severe conditions, enabling MoMA to accurately classify the case as a moderate injury. The lower section provides an example where clinical text and lab results are condensed into a concise summary by the specialist and aggregator agents. This summarized information allows the predictor agent to make predictions with improved interpretability compared to directly using all input text for classification.

Supplementary Material

Supplementary Note 1 Additional Evaluation Results

Supplementary Table 1 and Supplementary Table 2 present primary metrics (macro/micro-F1 for multiclass classification tasks, AUROC and AUPR for binary classification tasks). Supplementary Table 3 shows alternative metrics (macro-AUROC for multiclass classification tasks, F1 for binary classification tasks). Supplementary Table 4, Supplementary Table 5, Supplementary Table 6 and Supplementary Table 7 present subgroup analyses of MoMA and baselines. Supplementary Table 8 and Supplementary Table 9 show ablation studies evaluating when removing non-text components from MoMA. Supplementary Table 10 present the performance comparison of the published baselines (1D-CNN) for unhealthy alcohol use screening trained with different sample sizes. Supplementary Table 11 compares the results when adding a third modality (lab measurements) to MoMA in the chest trauma and multitask trauma severity stratification tasks.

Supplementary Table 1: Discrimination performance on the test set for chest trauma severity stratification and unhealthy alcohol use screening

	Chest trauma severity stratification		*Unhealthy alcohol use screening*	
	macro-F1	micro-F1	AUROC	AUPR
Published baseline	0.636 (0.609,0.663)	0.692 (0.666,0.717)	0.714 (0.655,0.771)	0.431 (0.339,0.517)
Cross-attention fusion	0.632 (0.601,0.663)	0.770 (0.747,0.793)	0.647 (0.581,0.709)	0.339 (0.255,0.424)
MoE fusion	0.630 (0.605,0.654)	0.756 (0.730,0.785)	0.670 (0.610,0.729)	0.374 (0.284,0.467)
LLaVA-Med	0.802 (0.775,0.829)	0.883 (0.868,0.899)	-	-
MoMA	0.834 (0.806,0.861)	0.903 (0.886,0.920)	0.755 (0.702,0.807)	0.491 (0.398,0.580)

Supplementary Table 2: Discrimination performance on the test set for multitask chest and spine trauma severity stratification

	Multitask Chest		*Multitask Spine*	
	macro-F1	micro-F1	macro-F1	micro-F1
Published baseline	0.630 (0.599,0.662)	0.794 (0.771,0.816)	0.494 (0.476,0.513)	0.781 (0.757,0.803)
Cross-attention fusion	0.621 (0.585,0.648)	0.784 (0.759,0.806)	0.495 (0.476,0.512)	0.783 (0.761,0.805)
MoE fusion	0.609 (0.581,0.647)	0.777 (0.750,0.803)	0.486 (0.464,0.511)	0.775 (0.749,0.801)
LLaVA-Med	0.806 (0.773,0.832)	0.889 (0.870,0.906)	0.694 (0.653,0.735)	0.856 (0.837,0.876)
MoMA	0.845 (0.817,0.871)	0.913 (0.897,0.928)	0.761 (0.721,0.800)	0.883 (0.864,0.900)

Supplementary Table 3: Alternative metrics

	Chest trauma severity stratification	*Unhealthy alcohol use screening*	*Multitask trauma severity stratification*	
	macro-AUROC	F1	macro-AUROC (chest)	macro-AUROC (spine)
Published baseline	0.826 (0.806,0.844)	0.400 (0.332,0.474)	0.865 (0.847,0.883)	0.844 (0.822,0.864)
Cross-attention fusion	0.815 (0.797,0.831)	0.403 (0.344,0.478)	0.851 (0.830,0.869)	0.836 (0.807,0.851)
MoE fusion	0.814 (0.800,0.829)	0.423 (0.364,0.503)	0.849 (0.827,0.866)	0.832 (0.804,0.846)
LLaVA-Med	0.953 (0.937,0.968)	-	0.956 (0.940,0.969)	0.910 (0.894,0.935)
MoMA	0.960 (0.945,0.973)	0.478 (0.385,0.558)	0.958 (0.947,0.968)	0.935 (0.920,0.948)

Supplementary Table 4: Subgroup Analysis for Chest Trauma Severity Stratification

	macro-F1				*micro-F1*			
	Male	Female	White	Non-white	Male	Female	White	Non-white
Published baseline	0.599 (0.564,0.634)	0.696 (0.564,0.634)	0.638 (0.610,0.667)	0.612 (0.524,0.702)	0.646 (0.612,0.681)	0.760 (0.723,0.798)	0.695 (0.668,0.722)	0.656 (0.571,0.738)
Cross-attention fusion	0.622 (0.583,0.661)	0.642 (0.590,0.694)	0.636 (0.604,0.669)	0.586 (0.496,0.680)	0.755 (0.725,0.786)	0.792 (0.754,0.829)	0.777 (0.752,0.801)	0.703 (0.619,0.786)
MoE fusion	0.625 (0.575,0.658)	0.638 (0.601,0.672)	0.631 (0.610,0.653)	0.570 (0.509,0.640)	0.731 (0.709,0.754)	0.779 (0.748,0.810)	0.766 (0.741,0.790)	0.702 (0.647,0.755)
LLaVA-Med	0.800 (0.762,0.837)	0.819 (0.769,0.863)	0.812 (0.780,0.843)	0.761 (0.666,0.854)	0.878 (0.853,0.901)	0.908 (0.882,0.934)	0.895 (0.877,0.913)	0.833 (0.762,0.893)
MoMA	0.828 (0.790,0.864)	0.845 (0.799,0.888)	0.841 (0.810,0.870)	0.819 (0.756,0.868)	0.893 (0.870,0.916)	0.919 (0.893,0.942)	0.910 (0.892,0.926)	0.886 (0.841,0.934)

Supplementary Table 5: Subgroup Analysis for **Multitask** Chest Trauma Severity Stratification

	macro-F1				*micro-F1*			
	Male	Female	White	Non-white	Male	Female	White	Non-white
Published baseline	0.605 (0.568,0.644)	0.667 (0.615,0.720)	0.629 (0.595,0.663)	0.621 (0.528,0.717)	0.757 (0.727,0.786)	0.849 (0.818,0.883)	0.794 (0.770,0.817)	0.797 (0.726,0.869)
Cross-attention fusion	0.571 (0.528,0.600)	0.611 (0.565,0.659)	0.585 (0.545,0.621)	0.579 (0.502,0.655)	0.743 (0.714,0.771)	0.839 (0.799,0.878)	0.783 (0.742,0.820)	0.781 (0.709,0.867)
MoEfusion	0.575 (0.531,0.617)	0.610 (0.568,0.673)	0.584 (0.546,0.622)	0.578 (0.505,0.660)	0.760 (0.731,0.789)	0.855 (0.821,0.886)	0.800 (0.775,0.826)	0.785 (0.701,0.862)
LLaVA-Med	0.784 (0.745,0.823)	0.833 (0.785,0.879)	0.807 (0.775,0.838)	0.763 (0.668,0.853)	0.870 (0.844,0.895)	0.916 (0.890,0.939)	0.893 (0.874,0.911)	0.844 (0.774,0.905)
MoMA	0.851 (0.816,0.884)	0.834 (0.788,0.879)	0.840 (0.810,0.869)	0.861 (0.785,0.931)	0.908 (0.887,0.929)	0.919 (0.896,0.942)	0.915 (0.898,0.930)	0.892 (0.833,0.940)

Supplementary Table 6: Subgroup Analysis for **Multitask** Spine Trauma Severity Stratification

	macro-F1				*micro-F1*			
	Male	Female	White	Non-white	Male	Female	White	Non-white
Published baseline	0.475 (0.451,0.498)	0.528 (0.498,0.557)	0.498 (0.479,0.517)	0.452 (0.380,0.520)	0.758 (0.727,0.788)	0.815 (0.780,0.850)	0.784 (0.758,0.808)	0.750 (0.679,0.821)
Cross-attention fusion	0.471 (0.445,0.499)	0.472 (0.441,0.504)	0.485 (0.456,0.517)	0.430 (0.341,0.509)	0.766 (0.731,0.799)	0.816 (0.782,0.847)	0.785 (0.760,0.805)	0.755 (0.683,0.826)
MoEfusion	0.480 (0.460,0.501)	0.519 (0.490,0.552)	0.492 (0.473,0.511)	0.443 (0.351,0.524)	0.755 (0.721,0.781)	0.820 (0.779,0.858)	0.781 (0.750,0.809)	0.750 (0.674,0.823)
LLaVA-Med	0.705 (0.656,0.753)	0.652 (0.574,0.729)	0.704 (0.660,0.747)	0.571 (0.435,0.733)	0.839 (0.813,0.865)	0.881 (0.853,0.910)	0.859 (0.838,0.879)	0.832 (0.762,0.893)
MoMA	0.792 (0.747,0.835)	0.749 (0.672,0.828)	0.762 (0.720,0.802)	0.738 (0.568,0.878)	0.878 (0.855,0.901)	0.890 (0.860,0.916)	0.882 (0.863,0.899)	0.874 (0.820,0.931)

Supplementary Table 7: Subgroup Analysis for Unhealthy Alcohol Use Screening

	AUROC		*AUPR*	
	Male	Female	Male	Female
Published baseline	0.733 (0.661,0.803)	0.682 (0.583,0.774)	0.491 (0.376,0.597)	0.346 (0.210,0.479)
Cross-attention fusion	0.681 (0.602,0.757)	0.585 (0.476,0.691)	0.378 (0.276,0.490)	0.329 (0.191,0.471)
MoE fusion	0.667 (0.590,0.742)	0.674 (0.571,0.770)	0.398 (0.284,0.516)	0.364 (0.220,0.515)
MoMA	0.782 (0.719,0.839)	0.745 (0.659,0.838)	0.543 (0.429,0.648)	0.468 (0.316,0.616)

Supplementary Table 8: Ablation study

	Chest trauma severity stratification		*Unhealthy alcohol use screening*	
	macro-F1	micro-F1	AUROC	AUPR
Non-text modality removed	0.778 (0.745,0.809)	0.858 (0.814,0.900)	0.736 (0.690,0.781)	0.470 (0.387,0.549)
MoMA	0.834 (0.806,0.861)	0.903 (0.886,0.920)	0.755 (0.702,0.807)	0.491 (0.398,0.580)

Supplementary Table 9: Ablation study

	Multitask Chest		*Multitask Spine*	
	macro-F1	micro-F1	macro-F1	micro-F1
Non-text modality removed	0.797 (0.767,0.828)	0.892 (0.875,0.909)	0.736 (0.696,0.779)	0.878 (0.860,0.897)
MoMA	0.845 (0.817,0.871)	0.913 (0.897,0.928)	0.761 (0.721,0.800)	0.883 (0.864,0.900)

Supplementary Table 10: Comparison of the published baselines (1D-CNN) for unhealthy alcohol use screening using different training data

	AUROC	AUPR
Published model (sample size: 54,915)	0.714 (0.655,0.771)	0.431 (0.339,0.517)
Retrained model (sample size: 1,576)	0.641 (0.566,0.719)	0.325 (0.231,0.419)

Supplementary Table 11: Comparison of MoMA with two (text+radiographs) and three modalities (text+radiographs+labs) combined for chest trauma and multitask trauma severity stratification

	Chest Trauma		*Multitask Chest*		*Multitask Spine*	
	macro-F1	micro-F1	macro-F1	micro-F1	macro-F1	micro-F1
Two Modalities (text+radiographs)	0.834 (0.806,0.861)	0.903 (0.886,0.920)	0.845 (0.817,0.871)	0.913 (0.897,0.928)	0.761 (0.721,0.800)	0.883 (0.864,0.900)
Three Modalities (text+radiographs+labs)	0.836 (0.805,0.863)	0.899 (0.885,0.922)	0.841 (0.815,0.869)	0.909 (0.893,0.925)	0.765 (0.724,0.805)	0.891 (0.869,0.904)

Supplementary Note 2 Optimization Hyperparameters

Supplementary Table 12, Supplementary Table 13, and Supplementary Table 14 are the technical details of developing MoMA for the three evaluation tasks. Full scripts and configs are available in our GitLab repository at `https://git.doit.wisc.edu/smph-public/dom/uw-icu-data-science-lab-public/moma`

Supplementary Table 12: Default hyperparameters for chest trauma severity stratification

Parameter	**Value**
Specialist agent	CXR-LLAVA-v2
Aggergator agent	Llama-3 8B
Predictor agent	Llama-3 8B
Optimizer	AdamW
Learning rate	1×10^{-4}
Load in 8bit	True
LoRA alpha	16
LoRA dropout	0.05
LoRA rank	128
Batch size	2
Max steps	4500
Warm-up steps	2
Gradient accumulation steps	1
Loss function	categorical cross-entropy

Supplementary Table 13: Default hyperparameters for multitask chest and spine trauma severity stratification

Parameter	**Value**
Specialist agent	CXR-LLAVA-v2
Aggergator agent	Llama-3 8B
Predictor agent	Llama-3 8B
Optimizer	AdamW
Learning rate	1×10^{-4}
Load in 8bit	True
LoRA alpha	16
LoRA dropout	0.05
LoRA rank	128
Batch size	2
Max steps	3500
Warm-up steps	2
Gradient accumulation steps	1
Loss function	Total categorical cross-entropy of the two sub-tasks

Supplementary Table 14: Default hyperparameters for unhealthy alcohol use screening

Parameter	Value
Specialist agent	Llama-3 8B
Aggergator agent	Llama-3 8B
Predictor agent	Llama-3 8B
Optimizer	AdamW
Learning rate	1×10^{-4}
Load in 8bit	True
LoRA alpha	32
LoRA dropout	0.05
LoRA rank	32
Batch size	2
Max steps	4500
Weight decay	0.01
Warm-up steps	2
Gradient accumulation steps	1
Loss function	binary cross-entropy with logits

Supplementary Note 3 Prompts

Supplementary Note 3.1 Prompt for the chest X-ray specialist agent in the chest trauma severity stratification task

We did not provide prompts to CXR-LLAVA [1], the chest X-ray specialist agent, as it performs effectively without additional prompting.

Supplementary Note 3.2 Prompt for the aggregator agent in the chest trauma severity stratification task

```
As an experienced trauma physician, your task is to review the clinical notes, radiology
reports, and LLM-generated radiology reports. Write a summary focusing on identifying and
summarizing chest trauma injuries, and determine the chest Abbreviated Injury Scale (AIS).

Follow these steps to complete the task:

1. Extract and summarize information related to the severity of chest trauma from the provided
    clinical notes and radiology reports. Do not include injuries in body regions outside the chest.

2. If the reports of X-RAY CHEST AP/PA/Single VIEW are not available, summarize the LLM-generated
    radiology reports as complementary information. Ensure that the LLM-generated reports do
    not overwrite clinical notes if they contradict each other.

3. Based on your summary, determine the chest AIS score (ranging from 0 to 6).
    Ensure the assessment is exclusively based on trauma-related conditions/symptoms.

4. Provide your conclusion as a single-digit number ranging from 0 to 6.
```

Supplementary Note 3.3 Prompt for the aggregator agent in the multitask chest and spine trauma severity stratification

```
As an experienced trauma physician, your task is to review the clinical notes, radiology
reports, and LLM-generated radiology reports. Write a summary focusing on identifying and
summarizing chest and spine trauma injuries, and determine the chest Abbreviated Injury Scale (AIS).

Follow these steps to complete the task:

1. Extract and summarize information related to the severity of only chest
    and spine trauma from the provided
    clinical notes and radiology reports. Do not include injuries in body regions
    outside chest and spine.

2. If the reports of X-RAY CHEST AP/PA/Single VIEW are not available, summarize
    the LLM-generated radiology reports as complementary information. Ensure that
    the LLM-generated reports do not overwrite clinical notes if they contradict each other.

3. Based only on trauma-related conditions or symptoms, assign an Abbreviated
    Injury Scale (AIS) score (0-6) for each region.
    Remember that only conditions/symptoms caused by trauma injuries should be
    used to determine the AIS scores.

4. Based on the AIS scores of chest and spine injuries, translate to the Severity
    Category for chest and spine:
    AIS = 0 → Negative
    AIS = 1 or AIS = 2 → Moderate
    AIS > 2 → Serious
```

Supplementary Note 3.4 Prompt for the lab measurement specialist agent in the unhealthy alcohol use screening task

```
As an expert in screening for unhealthy alcohol use, carefully review the provided lab
measurements and generate a concise summary highlighting potential indicators of unhealthy
alcohol use based on your analysis. Let's think through this step by step:
1. Identify any initial measurements commonly linked to alcohol consumption or misuse like serum
blood alcohol levels.
2. Consider labs with indirect evidence for unhealthy alcohol consumption (e.g., elevated liver
enzymes, mean corpuscular volume).
3. Incorporate labs that have previously been shown to serve as biomarkers of unhealthy alcohol
use.

Make sure your response is short and concise. Avoid being verbose.

Here are a few examples:
Direct Indicator: The serum blood alcohol level of 12 mg/dL is above the legal
limit and indicates recent alcohol consumption.
Indirect Evidence: Elevated AST and ALT levels, along with an increased MCV, suggest liver
dysfunction and macrocytosis, both of which are commonly associated with chronic alcohol misuse.
Biomarker: The GGT level is significantly elevated, which can serve as a biomarker for
heavy alcohol use.
```

Supplementary Note 3.5 Prompt for the aggregator agent in the unhealthy alcohol use screening task

```
Role:
You are an alcohol screener working within a healthcare system, responsible for
determining whether a patient has exhibited signs of unhealthy alcohol use over the
past three months. Your evaluation will be based on clinical summaries generated by
LLM agents, which include clinical notes and lab measurements.

Objective:
Develop a focused summary of the patient's alcohol use. Ensure that no personally
identifiable information (PHI) is included.

Task Instructions:

1. Assess Evidence of Alcohol Misuse:
- Review the clinical summaries to identify any details related to alcohol use,
    including behavioral patterns, attempts to manage drinking, or external concerns.
```

```
- Focus on direct evidence and ensure the summary highlights key findings related to
    alcohol use while avoiding unnecessary or redundant information.

2. Summarize Lab Measurements:
- Pay close attention to direct evidence from lab results, such as blood alcohol
    concentration (BAC) levels, as they provide clear indications of alcohol use.
- Include other lab abnormalities only if they are explicitly connected to alcohol use.

3. Evaluate Causes of Lab Abnormalities:
- For any mentioned lab abnormalities, review the clinical summaries to determine if
    they may have causes unrelated to alcohol use.
- Exclude such lab results from the summary and explicitly state when a lab abnormality
    have an alternative cause.

4. Compose a Unified Summary:
- Write a comprehensive summary of the patient’s alcohol use, integrating relevant
    details from both the clinical summaries and lab results.
- Ensure the summary prioritizes key findings, focusing primarily on direct evidence
    such as BAC levels and behavioral indications of alcohol use.
```

Supplementary Note 3.6 Prompt for summarizing clinical notes for LLaVA-Med in the multitask chest and spine trauma severity stratification

```
You are a clinical summarization assistant.
Your job is to read the given ED notes and radiology reports, then extract only
the details related to chest trauma and spine trauma, separately.

1. Produce two labeled sections in your response:
 - Chest Trauma Summary:
 - Spine Trauma Summary:

2. Keep each summary short and self-contained. Do not mention or quote which
section(s) of the note the information came from.

3. If no chest trauma is mentioned, exactly reply:
> No chest trauma mentioned in the clinical note.
   If no spine trauma is mentioned, exactly reply:
> No spine trauma mentioned in the clinical note.
   If neither chest nor spine trauma is mentioned, exactly reply:
> No chest or spine trauma mentioned in the clinical note.

4. Do not include any additional commentary or information beyond the two summaries
or one of the exact ‘‘No ... mentioned’’ statements.
```

Supplementary Note 3.7 Prompt for summarizing clinical notes for LLaVA-Med in the chest trauma severity stratification

```
You are a clinical summarization assistant.
Your job is to read the given ED notes and radiology reports, then extract only the
details related to chest trauma.

1. Keep the summary short and self-contained. Do not mention or quote which
section(s) of the note the information came from.

3. If no chest trauma is mentioned, exactly reply:
> No chest trauma mentioned in the clinical note.

4. Do not include any additional commentary or information beyond the summary or the
```

```
exact ‘‘No ... mentioned’’ statements.
```

Supplementary Note 3.8 Prompt for generating severity predictions for LLaVA-Med in the chest trauma severity stratification

```
You are a radiology assistant specialized in chest trauma.
Given a chest X-ray and a brief clinical note summary,
classify the trauma severity on a scale from:
0 = no trauma
1 = minor or moderate trauma
2 = serious or greater than serious trauma
Reply with exactly one integer (like ’1,2’).
```

Supplementary Note 3.9 Prompt for generating severity predictions for LLaVA-Med in the multitask chest and spine trauma severity stratification

```
You are a radiology assistant specialized in chest and spine trauma.
Given a chest X-ray and a brief clinical note summary,
classify the trauma severity on a scale from:
0 = no trauma
1 = minor or moderate trauma
2 = serious or greater than serious trauma
Reply with exactly two integers separated by comma (like ’1,2’), one for chest
and one for spine, and no other text.
```

Supplementary Note 4 TRIPOD-LLM checklist

The TRIPOD-LLM [2] checklist is attached below. Based on the instructions, we selected “**LLM methods**" as the research design and “**Classification**" as the research task. All the relevant items are reported below in Supplementary Table 15

Section / Topic	*Checklist Item*	*Reported on Page*
Abstract		
Title	Identify the study as developing, fine-tuning, and/or evaluating the performance of an LLM, specifying the task, the target population, and the outcome to be predicted.	1
Objective	Specify the study objectives, including whether the study describes LLMs development, tuning, and/or evaluation	1
Methods	Describe the key elements of the study setting.	1
	Detail all data used in the study, specify data splits and any selective use of data.	1
	Specify the name and version of LLM used.	7
	Briefly summarize the LLM-building steps, including any fine-tuning, reward modeling, reinforcement learning with human feedback (RLHF), etc.	1
	Describe the specific tasks performed by the LLMs (e.g., medical QA, summarization, extraction), highlighting key inputs and outputs used in the final LLM.	1
	Specify the evaluation datasets/populations used, including the endpoint evaluated, and detail whether this information was held out during training/tuning where relevant, and what measure(s) were used to evaluate LLM performance.	1
Results	Give an overall report and interpretation of the main results.	2,3

Discussion	Explicitly state any broader implications or concerns that have arisen in light of these results.	2,3
Introduction		
Background	Explain the healthcare context / use case (e.g., administrative, diagnostic, therapeutic, clinical workflow) and rationale for developing or evaluating the LLM, including references to existing approaches and models.	1,2,3
Objectives	Specify the study objectives, including whether the study describes the initial development, fine-tuning, or validation of an LLM (or multiple stages).	1,2,3
Methods		
Data	Describe the sources of data separately for the training, tuning, and/or evaluation datasets and the rationale for using these data (e.g., web corpora, clinical research/trial data, EHR data).	3
	Describe the relevant data points and provide a quantitative and qualitative description of their distribution and other relevant descriptors of the dataset (e.g., source, languages, countries of origin)	3,15
	Specifically state the date of the oldest and newest item of text used in the development process (training, fine-tuning, reward modeling) and in the evaluation datasets.	3,15
	Describe any data pre-processing and quality checking, including whether this was similar across text corpora, institutions, and relevant sociodemographic groups.	3
	Describe how missing and imbalanced data were handled and provide reasons for omitting any data.	3
Analytical Methods	Report the LLM name, version, and last date of training or use during inference.	7
	Specify the type of LLM architecture, and LLM building steps, including any hyperparameter tuning (e.g., temperature, length limits, penalties), prompt engineering, and any inference settings (e.g., seed, temperature, max token length) as relevant.	6,7,8,9,10
	Report details of LLM development process from text input to outcome generation, such as training, fine-tuning procedures, and alignment strategy (e.g., reinforcement learning, direct preference optimization, etc.) and alignment goals (e.g., helpfulness, honesty, harmlessness, etc.).	6,7,8,9,10
	Specify the initial and post-processed output of the LLM (e.g., probabilities, classification, unstructured text).	6,7,8,9,10
	Provide details and rationale for any classification and how the probabilities were determined and thresholds identified.	3
LLM Output	If outcome assessment requires subjective interpretation, describe the qualifications of the assessors, any instructions provided, relevant information on demographics of the assessors, and inter-assessor agreement.	3
	Specify how performance was compared to other LLMs, humans, and other benchmarks or standards.	4,5,16,17,18
Annotation	If annotation was done, report how text was labeled, including providing specific annotation guidelines with examples.	3
	If annotation was done, report how many annotators labeled the dataset(s), including the proportion of data in each dataset that were annotated by more than 1 annotator.	3
	If annotation was done, provide information on the background and experience of the annotators, and the inter-annotator agreement.	3

Prompting	If research involved prompting LLMs, provide details on the processes used during prompt design, curation, and selection.	7,8,9
	If research involved prompting LLMs, report what data were used to develop the prompts.	7,8,9
Instruction Tuning / Alignment	If instruction tuning/alignment strategies were used, what were the instructions and interface used for evaluation, and what were the characteristics of the populations doing evaluation?	N/A. Instruction tuning/alignment strategies are not used.
Compute	Report compute, or proxies thereof (e.g., time on what and how many machines, cost on what and how many machines, inference time, floating-point operations per second (FLOPs)), required to carry out methods.	10
Ethics Approval	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent.	10
Open Science	Give the source of funding and the role of the funders for the present study.	14
	Declare any conflicts of interest and financial disclosures for all authors.	14
	Provide details of the availability of the study data.	14
	Provide details of the availability of the code to reproduce the study results.	14
Results		
Performance	Report LLM performance according to pre-specified metrics (see item 7a) and/or human evaluation (see item 7d).	3,4,15,16,17,18
LLM Updating	If applicable, report the results from any LLM updating, including the updated LLM and subsequent performance.	3,4,15,16,17,18
Discussion		
Interpretation	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies.	5,6
Limitations	Discuss any limitations of the study and their effects on any biases, statistical uncertainty, and generalizability.	6
Usability of the LLM in context	Discuss any next steps for future research, with a specific view to applicability and generalizability of the LLM.	6

Supplementary Table 15: TRIPOD-LLM checklist

References for Supplementary Material

[1] Seowoo Lee et al. “Cxr-llava: Multimodal large language model for interpreting chest x-ray images”. In: *arXiv preprint arXiv:2310.18341* (2023).

[2] Jack Gallifant et al. “The TRIPOD-LLM reporting guideline for studies using large language models”. In: *Nature Medicine* (2025), pp. 1–10.