OIDA-QA: A Multimodal Benchmark for Analyzing the Opioid Industry Documents Archive

Xuan Shen[1], Brian Wingenroth[2], Zichao Wang[3], Jason Kuen[3], Wanrong Zhu[3], Ruiyi Zhang[3], Yiwei Wang[4], Lichun Ma[5], Anqi Liu[2], Hongfu Liu[6], Tong Sun[3], Kevin S. Hawkins[2], Kate Tasker[7], G. Caleb Alexander[2], Jiuxiang Gu[3]*

[1]Northeastern University
[2]Johns Hopkins University
[3]Adobe Research
[4]University of California, Merced
[5]National Institutes of Health
[6]Brandeis University
[7]University of California, San Francisco
shen.xu@northeastern.edu, wingenroth@jhu.edu, {jackwa, kuen, wzhu, ruizhang, tsun}@adobe.com, yiweiwang2@ucmerced.edu, lichun.ma@nih.gov, aliu@cs.jhu.edu, hongfuliu@brandeis.edu, khawki29@jhu.edu, Kate.Tasker@ucsf.edu, galexan9@jhu.edu, jigu@adobe.com

Abstract

The opioid crisis represents a significant moment in public health that reveals systemic shortcomings across regulatory systems, healthcare practices, corporate governance, and public policy. Analyzing how these interconnected systems simultaneously failed to protect public health requires innovative analytic approaches for exploring the vast amounts of data and documents disclosed in the UCSF-JHU Opioid Industry Documents Archive (OIDA). The complexity, multimodal nature, and specialized characteristics of these healthcare-related legal and corporate documents necessitate more advanced methods and models tailored to specific data types and detailed annotations, ensuring the precision and professionalism in the analysis. In this paper, we tackle this challenge by organizing the original dataset according to document attributes and constructing a benchmark with 400k training documents and 10k for testing. From each document, we extract rich multimodal information—including textual content, visual elements, and layout structures—to capture a comprehensive range of features. Using multiple AI models, we then generate a large-scale dataset comprising 360k training QA pairs and 10k testing QA pairs. Building on this foundation, we develop domain-specific multimodal Large Language Models (LLMs) and explore the impact of multimodal inputs on task performance. To further enhance response accuracy, we incorporate historical QA pairs as contextual grounding for answering current queries. Additionally, we incorporate page references within the answers and introduce an importance-based page classifier, further improving the precision and relevance of the information provided. Preliminary results indicate the improvements with our AI assistant in document information extraction and question-answering tasks.

Datasets —
https://huggingface.co/datasets/opioidarchive/oida-qa

*Corresponding author: Jiuxiang Gu (jigu@adobe.com)

Introduction

The opioid crisis has severely impacted global health, exposing weaknesses in healthcare systems and contributing to issues like domestic violence and child abuse (NIDA 2024; Oderda et al. 2015; Swedo et al. 2020). In 2019, 10.1 million Americans reported opioid misuse, and from June 2021 to May 2022, opioids were involved in 90% of the 108,000 U.S. overdose deaths (CDC 2020; Tanz et al. 2022). Though effective for pain relief, opioids can induce euphoria, resulting in misuse and addiction, particularly in regions with inadequate healthcare services, underscoring the urgent need to address drug misuse (Birnbaum et al. 2011). Recent advances in AI have increasingly revealed its potential in healthcare. Particularly, AI assistants based on LLMs have emerged as powerful tools for extracting insights from unstructured medical data and assisting healthcare professionals in clinical Question Answering (QA) (Lee et al. 2020; Singhal et al. 2023a; Lu, Dou, and Nguyen 2022). Thus, we wish to develop AI-driven solutions to address the opioid crisis with the public data from the UCSF-JHU Opioid Industry Documents Archive (OIDA 2021), a massive corpus of internal correspondence and other documents arising from the opioid industry.

While LLMs perform well on general QA tasks, the multimodal and long-context nature of healthcare data poses challenges such as complex reasoning and increased hallucination risks (Gao et al. 2023). Most LLMs also struggle with multi-turn interactions, limiting their ability to handle sequential user queries effectively. Furthermore, they often fail to link answers to specific pages or sections within documents, resulting in poor answer provenance and reduced trustworthiness. These challenges are compounded by the escalating opioid crisis and the rapid growth of the OIDA dataset. Thus, creating a low-cost, reliable, and scalable multimodal LLM accessible to the public is essential.

In this paper, we propose OIDA-QA, a multimodel multi-

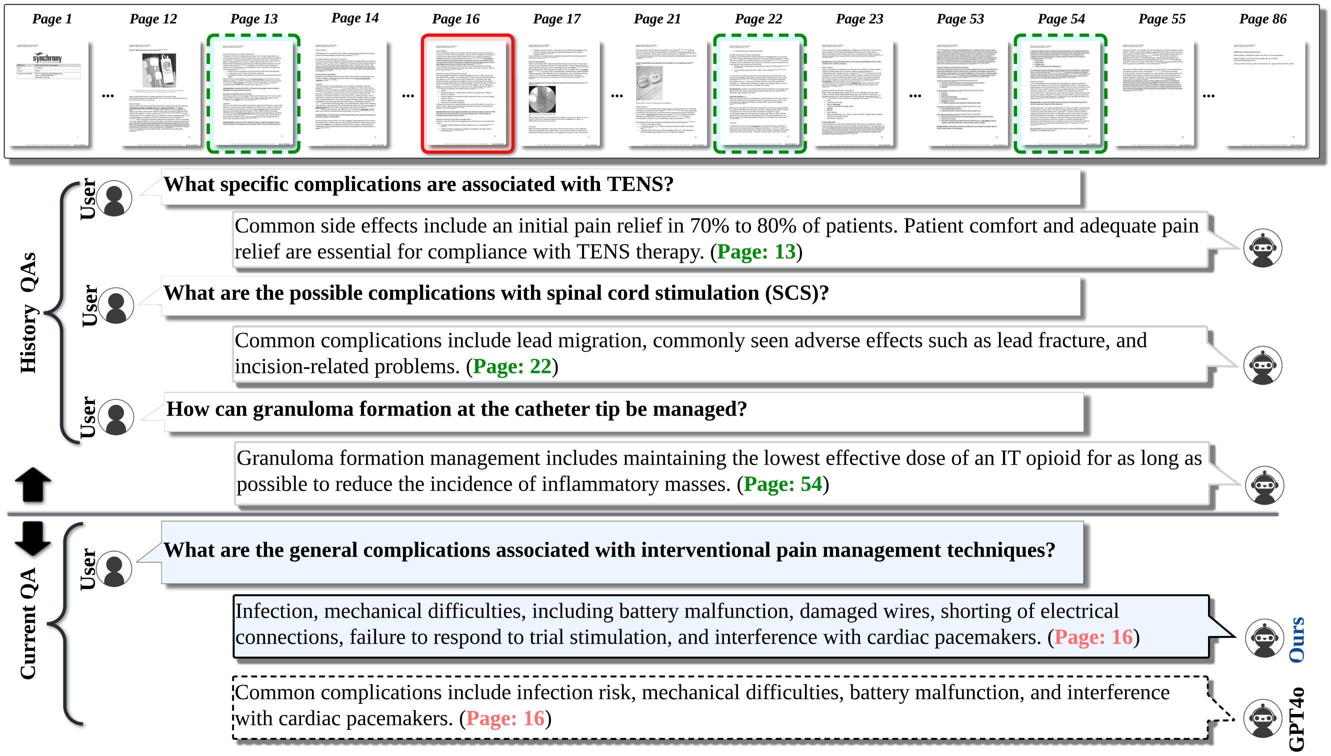


Figure 1: Overview of the proposed OIDA-QA: Our QA dataset enables multi-hop QA with answers grounded to specific pages, effectively handling multi-page documents by extracting dense information (textual, visual, and layout) from scanned documents.

page and multi-hop document question-answering benchmark based on the OIDA. To effectively handle large-scale document data, we start by analyzing its distribution using the taxonomy proposed in ADOPD (Gu et al. 2024), combined with the CLIP (Radford et al. 2021) model finetuned on ADOPD's image-caption pairs. Utilizing the taxonomy-derived clusters, we identify the 20 clusters with the most documents and further diversify based on sub-categories and page count. We compile 20K PDF documents from each selected cluster to create the final training set. To develop a comprehensive understanding of each PDF document, we enrich the original PDFs by extracting textual information (OCR words), visual elements (tags and masks), and layout details (bounding boxes). These enriched, model-assisted multimodal annotations provide the foundation for generating QA pairs. Then, we introduce user personas (Chan et al. 2024) when generating questions, simulating users from various backgrounds to create a diverse set of queries. Meanwhile, answerability and page-level grounding of the questions are ensured by leveraging a multimodal LLM, which serves as both a verifier and locator for the generated QA pairs. Furthermore, **we recruit medical professionals, including doctors and nurses, to annotate and refine 100k QA pairs in our dataset, ensuring the accuracy and reliability of the majority of the generated QA pairs**.

To handle long-context reasoning, we enable LLMs to learn relations across extended contexts and ground answers to the right pages using instruction-based page-finding prompts and a key-page identification model. By expanding the model's context window during training, we capture the semantic information of long documents. During testing, for documents that exceed the model's processing capacity, we use the key page identification model to preliminarily locate important pages and feed these into the model to obtain precise answers and their corresponding page locations. This approach effectively addresses the challenge of processing ultra-long texts, improving both efficiency and accuracy. The overview of our proposed method is visualized in Figure 1. Experiments show that our model performs excellently in both answering questions and locating pages, validating the effectiveness of our approach. This highlights the potential of multi-turn dialogue systems in the medical domain. We believe our system plays a vital role in medical information retrieval and opioid abuse prevention, providing robust technical support to address this pressing public health crisis.

The contributions of this paper are summarized as follows:

- We introduce OIDA-QA, a multimodal document QA benchmark based on the OIDA, along with an effective method for enriching PDF documents with textual, visual, and layout annotations. This provides a data foundation for the exploration of AI-driven solutions to the opioid crisis.
- We utilize the dense extracted data and LLMs to create a persona-based, long-context, multi-page, multi-hop QA benchmark. By integrating user personas and page-level references, OIDA-QA benefits to a broader range of users and enhances the answer verifiability of the benchmark.

- We develop a scalable model system for long-context processing and answer page grounding. By employing instruction-based prompts and introducing a key page identification model, our approach enhances the model's ability to locate relevant pages within lengthy documents, improving both efficiency and accuracy.
- We validate the effectiveness of our proposed methods through extensive experiments. The results demonstrate strong performance in both question answering and page locating tasks, confirming the efficacy of our approach.

Related Work

Document Understanding Pre-training large models that handle both textual and visual information have proven highly effective for document understanding tasks (Huang et al. 2022; Da et al. 2023; Shen et al. 2024b,c,d, 2025a,c). Unlike traditional LLMs that process plain text, document understanding requires models to consider layout information (Tu et al. 2023; Yu et al. 2023; Shen et al. 2025b, 2024a, 2022, 2025e,d,g). Recently, LLMs and Multimodal LLMs (MLLMs) (OpenAI 2022, 2023; Yang et al. 2023; Liu et al. 2025a,b) have demonstrated outstanding zero-shot performance across a wide range of Natural Language Processing (NLP) and Computer Vision (CV) tasks. Leveraging LLMs for zero-shot document understanding has also shown promise (Perot et al. 2023; Zhang et al. 2023; Shen et al. 2023, 2025f). Although existing models have demonstrated promising results in document tasks, most training data originates from traditional datasets that lack sufficient document data. The OIDA data poses challenges for existing models due to its multi-page, multimodal characteristics. However, previous works focus on general vision-language domain, not for multi-page document understanding tasks.

Healthcare Question Answering With over one-third of American adults seeking medical information online (Fox and Duggan 2012), there is a growing need for robust healthcare QA systems to support patient consultations. Previous healthcare QA systems with Pre-trained Language Models (Pergola et al. 2021) faced challenges in real-world applications due to limited language understanding and generation capabilities (Liu et al. 2023). In healthcare, LLMs like Med-PaLM 2 (Singhal et al. 2023b) have shown remarkable progress, achieving 86.5% on the USMLE dataset and outperforming earlier models (Singhal et al. 2022). This advancement extends to other medical datasets, including MedMCQA (Pal, Umapathi, and Sankarasubbu 2022) and PubMedQA (Jin et al. 2019). However, as summarized in Table 1, existing healthcare QA datasets often lack multi-turn conversations and grounding information. Only datasets such as MedDialog-CN (He et al. 2020) and MedDialog-EN (He et al. 2020) provide multi-round dialogues but without grounding.

OIDA-QA Benchmark

Data Collection and Extraction

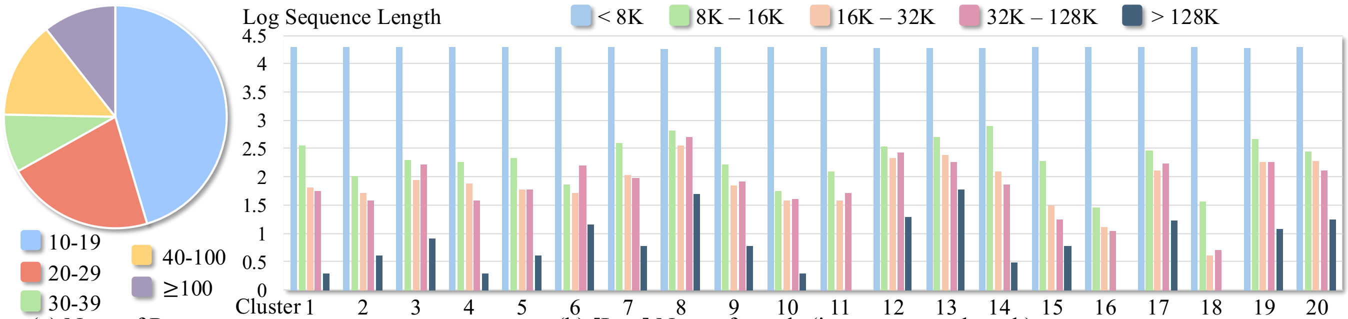
The original OIDA data is available in PDF format[1], along with metadata, and extracted text for all documents. Layout and visual information are not available due to the limitations of OCR and document extraction models. For the OIDA documents, extracting as much detailed information as possible is essential to fully explore the model's capabilities and better serve the community. To fully leverage OIDA for developing an AI system, our data collection process includes the following steps: (1) analyzing the distribution of the OIDA dataset, (2) performing balanced data sampling, and (3) extracting multimodal information.

Data Distribution Analysis To understand the distribution of OIDA, we utilize the pre-trained CLIP model from ADOPD and its taxonomy to tag the first page of each document in a zero-shot manner. This is achieved by calculating the similarity between the visual features of the page images and the textual embeddings. The textual embeddings are composed of the proposed taxonomy labels from ADOPD, formatted as '*a photo of ¡candidate label¿*'. The top five most relevant labels are selected for document grouping, which categorizes each document into clusters based on the hierarchically structured ADOPD taxonomy. The predicted document tags and clustered results for all PDF documents offer a comprehensive overview of the OIDA dataset's distribution. This cluster-based sampling approach minimizes selection bias, resulting in a more diverse training set.

Data Sampling Let the entire dataset be denoted by $\mathcal{D}_{\text{full}}$. We downsample $\mathcal{D}_{\text{full}}$ by selecting the top K largest clusters based on distribution analysis above. For each cluster $k \in \{1, 2, \ldots, K\}$, we define the subset $\mathcal{D}_k = \{D_{k,1}, D_{k,2}, \ldots, D_{k,N_k}\}$, where each document $D_{k,i}$ consists of multiple pages. In our experiments, we set K to 20. For the sampling of each cluster, we balance the subcategories according to the labels and the number of pages. We collect a diverse training set from these 20 clusters, including 20K PDF documents per cluster (for a total of 400K documents). For the test set, we gather 500 PDF documents per cluster, resulting in a total of 10K documents. To construct multi-hop QA pairs, we select long documents—those exceeding 10 pages—from the extracted dataset for better QA generation. In Figure 2(a), we visualize the distribution of documents across eight different page counts within each cluster. Figure 2(b) visualizes the distribution of documents based on the logarithm of sequence length, emphasizing the long-sequence characteristics of OIDA-QA.

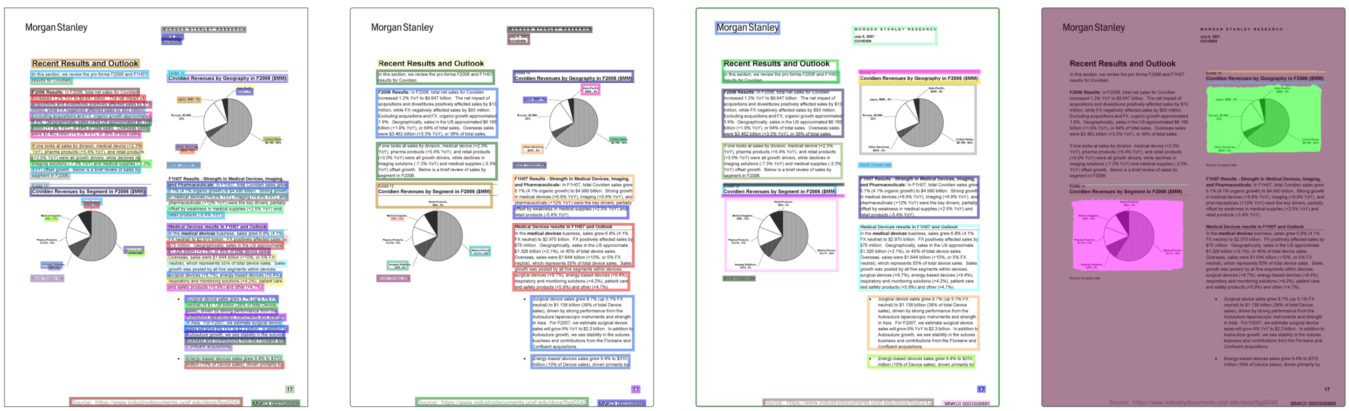
Document Data Extraction For each document $D_{k,i}$, we extract textual, visual, and layout information from each page. Using an OCR tool, we extract words to obtain a set of text lines, which are then grouped into paragraphs $\{\mathbf{p}_{k,i,j}\}$ using heuristic rules. We assign each paragraph $\mathbf{p}_{k,i,j}$ a location $\mathbf{l}_{k,i,j} = \left(p_{k,i,j}, b_x^l, b_y^t, b_x^r, b_y^b\right)$, where $p_{k,i,j}$ is the page number, and $(b_x^l, b_y^t, b_x^r, b_y^b)$ are the normalized bounding box coordinates. Figure 3 illustrates that text lines alone do not capture the semantic relationships between words, and that merging words using rule-based methods is also limited by heuristic constraints. To address these limitations, we utilize the Doc2Box model (Gu et al. 2024) to extract text blocks that better preserve semantic structures. For visual information, we provide two outputs: CLIP tags and entity masks. The CLIP tags capture the high-level attributes of the documents.

[1] s3://opioid-industry-documents-archive-dataset-bucket/



(a) Num of Pages (b) [Log] Num of words (i.e., sequence length)

Figure 2: Data distribution of (a) number of pages and (b) number of words (i.e., sequence length).



(a) OCR Text Lines (b) Merged OCR (c) Doc2Box (d) Doc2Mask

Figure 3: Comparison of information extraction methods: (a) OCR text lines, (b) statistical merging results, (c) text blocks from Doc2Box, and (d) entity masks from Doc2Mask.

We apply the trained Doc2Mask model (Gu et al. 2024) to identify the entity masks. By combining textual, visual, and layout information, we create a comprehensive representation of each document that supports advanced processing tasks such as QA and information extraction.

Persona-Based Multi-Hop Question-Answering Dataset Generation

Persona Setup To simulate diverse user interactions and generate questions from various perspectives, we incorporate the vast Persona Hub (Chan et al. 2024), including over one billion personas, into our benchmark for question generation. We begin by generating the relevant personas for our benchmark. For each cluster, we randomly sample 500 personas from the full Persona Hub and employ the GPT-4o (OpenAI 2023) to generate 48 detailed personas in average based on the assigned labels of the cluster. The detailed personas, with attributes: *Name*, *Age*, *Gender*, *Major Background*, *Previous Experience*, and *Hobbies*, ensures the subsequent question generation process draws from a diverse user profiles.

Multi-Hop QA Generation We employ a model-assisted approach for question-answer data generation, structuring the process into question generation and answer generation using GPT-4o. Algorithm 1 detailed in Appendix illustrates the QA data generation process. Building upon our data extraction method described in Section , we use only the grouped text lines as input for this step. To ensure a wide range of perspectives in the question generation process, we introduce the characterized personas discussed earlier. For generating QA pairs, we employ GPT-4o instances in the following role: (1) a question generator that creates questions based on the document content and persona attributes; (2) an answer generator that determines if the answer is answerable and provides the corresponding response with referred page number simultaneously. (3) a QA pair decomposer that decomposes single QA into more than one QA sequence to form a coherent multi-turn conversation. As a result, we generate over 360k QA pairs with grounding information, indicating the specific page each answer refers to right side of Figure 4 shows the distribution of QA counts in our benchmark.

Method

In this section, we detail the training process of our model for multi-round QA over long documents with page-grounded answers. Our approach includes instruction tuning of the LLM and introduces a page-finding mechanism, implemented by adjusting the training loss to improve page grounding and proposing a page-finding model to efficiently locate relevant

Task	Source	Size	# Round	Grounding
General	Alpaca	52k	1	No
	Dolly	15k	1	No
	ShareGPT	48k	1	No
	DocVQA	50k	1	No
HealthTech	HealthCareMagic	100k	1	No
	iCliniq	7k	1	No
	MedDialog-CN	792k	≥1	No
	MedDialog-EN	257k	≥1	No
	MedInstruct	52k	1	No
	Medical Flash Cards	34k	1	No
	WikiDocPatient	5k	1	No
	PubMedQA	211k	1	No
	OIDA-QA (Ours)	**370k**	**≥2**	**Yes**

Table 1: Comparison of QA datasets across general QA and HealthTech domains, including biomedical and clinical.

pages within the documents.

Instruction Tuning for Multi-Hop QA

We fine-tune the pretrained LLM using instruction tuning to perform multi-hop QA tasks. Given the LLM as model with parameters θ, the model is trained to generate appropriate answers given an input context, a sequence of questions, and the dialogue history.

Let $\mathcal{D} = \{(C_i, H_i)\}_{i=1}^{N}$ denote the training dataset, where C_i is the input context for the i-th example, formulated based on the extracted sentences $\mathcal{S}$ and their locations; and $H_i = \left[\left(q_i^k, a_i^k\right)\right]_{k=1}^{K_i}$ is the dialogue history consisting of question-answer pairs, with K_i being the number of dialogue turns in the i-th example. At each turn j, the model generates the answer a_i^j conditioned on the context C_i, the current question q_i^j, and the previous dialogue history $H_i^{<j} = \left[\left(q_i^k, a_i^k\right)\right]_{k=1}^{j-1}$; the conditional probability of generating a_i^j is given by $P\left(a_i^j \mid C_i,\, q_i^j,\, H_i^{<j};\, \theta\right)$, where θ represents the model parameters. Finally, the training objective is to minimize the negative log-likelihood over the dataset as follows,

$$\mathcal{L}_{\text{QA}}(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{K_i} \log P\left(a_i^j \mid C_i,\, q_i^j,\, H_i^{<j};\, \theta\right). \quad (1)$$

This loss function encourages the model to generate accurate answers based on the context and dialogue history in a multi-turn setting.

Query-Based Page Finding

When dealing with long documents, the input context C_i may exceed the model's maximum context window size L_{max}. To address this, we truncate the context to fit within the allowable length, typically by focusing on the ground-truth page. However, this introduces a training-testing mismatch, as the model relies on the ground-truth page during training but lacks this information during testing. To mitigate this, we propose two improvements: (1) *Enhancing Page-Finding via Ground-Truth Content Reiteration* to improve content localization, and (2) *Training a Query-based Page Finder* to enable effective page identification without relying on ground-truth information during testing.

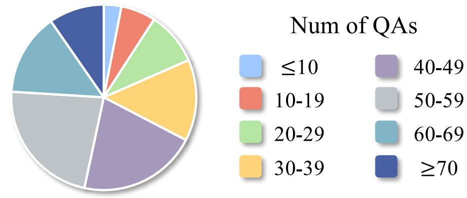


Figure 4: Visualization of the QA pair counts within our benchmark.

Enhancing Page-Finding via Content Reiteration The training objective in Section focuses on optimizing the model's ability to generate accurate answers based on the input context and questions. To enhance the model's *page-finding capability through content reiteration*, we introduce an additional optimization step. Specifically, we modify the model so that it outputs the associated page references along with a relevant portion of the context as a single, unified response. This is achieved by designing prompts that instruct the model to include page indices and pertinent context excerpts in its output, thereby reinforcing the link between the content and its location within the document. To formalize this enhancement, we adjust the conditional probability to jointly model the page indices and the relevant context portion as one output as follows,

$$P\left(p_i^j,\, \tilde{C}_i^j \mid C_i,\, q_i^j,\, H_i^{<j};\, \theta\right), \quad (2)$$

where p_i^j represents the page indices relevant to the query, and $\tilde{C}_i^j$ is the extracted portion of the context associated with those pages.

Additionally, we further employ a page-finding loss function that captures the joint probability of generating the correct page and context to balance the learning of both tasks as follows,

$$\mathcal{L}_{\text{PF}}(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{K_i} \log P\left(p_i^j,\, \tilde{C}_i^j \mid C_i,\, q_i^j,\, H_i^{<j};\, \theta\right). \quad (3)$$

During training, we incorporate both the QA loss $\mathcal{L}_{\text{QA}}(\theta)$ and the page-finding loss $\mathcal{L}_{\text{PF}}(\theta)$ to enable accurate answer generation and effective page-level grounding over long documents. We combine page-finding and QA data in a specific proportion to train the model on both tasks. This approach helps the model focus on relevant sections, output page references, and generate responses grounded in specific document parts, enhancing performance in multi-hop QA tasks.

Page Finder for Ultra-Long Contexts As mentioned earlier, during testing, the input sequence can become very long, potentially exceeding hardware limitations and causing significant delays. This motivates us to propose a method to assist in finding important pages during testing. Thus, we develop a separate *Page Finder* module to identify relevant pages before generating answers.

For multi-page documents, we first extract the content of each page. We use an encoder model based on Sentence Transformers (Reimers and Gurevych 2019) to generate embeddings for each page c_l and the query q_i^j as follows,

$$\mathbf{h}_l = \phi(c_l), \quad \mathbf{h}_q = \phi(q_i^j), \tag{4}$$

where ϕ denotes the Sentence Transformer encoder.

To train the Page Finder module, we utilize the Multiple Negatives Ranking Loss (Henderson et al. 2017), which is effective for training retrieval models with in-batch negatives. Given a batch of B query-document pairs $\{(q_b, c_b^+)\}_{b=1}^B$, where each q_b is a query and c_b^+ is its corresponding positive (relevant) page, we compute the similarity scores between each query and all document embeddings in the batch. Then, the total loss over the batch is then computed as follows,

$$\mathcal{L}_{\text{MNRL}} = -\frac{1}{B} \sum_{b=1}^{B} \log \frac{\exp(s_{b,b}/\tau)}{\sum_{k=1}^{B} \exp(s_{b,k}/\tau)}, \tag{5}$$

where $s_{b,k} = \cos(\mathbf{h}_{q_b}, \mathbf{h}_{c_k^+})$ is the similarity score between query q_b and document c_k^+, and τ is a temperature hyperparameter that controls the softness of the probability distribution. This loss encourages the model to assign high similarity scores to the correct query-document pairs while minimizing the scores for mismatched pairs within the batch, which serve as negative samples.

During inference, we employ the trained Page Finder to compute the relevance scores between the query and each page using cosine similarity. We then select the top-ranked pages based on these scores and extend the selection by adding additional adjacent pages until the context length limit L_{max} is reached as follows,

$$C_i^{\text{reduced}} = [c_{l_1}, c_{l_2}, \ldots, c_{l_K}], \tag{6}$$

where K is determined such that the combined length of the selected pages does not exceed L_{max}.

We then provide C_i^{reduced} to LLMs for answer generation as follows,

$$P\left(a_i^j \mid C_i^{\text{reduced}}, q_i^j, H_i^{<j}; \theta\right). \tag{7}$$

By integrating the Page Finder module into our system, we effectively handle ultra-long documents by focusing on the most relevant sections, mitigating hardware limitations and reducing inference delay. Meanwhile, training the Page Finder enables it to accurately retrieve the pages most pertinent to the query, enhancing the overall performance.

Experiments

Experimental Analysis

Training Details In our main experiments, we utilize the Mistral-7B-Instruct-v0.2 model (Jiang et al. 2023), employing full-parameter fine-tuning for improved optimization. The training is conducted over one epoch with a local batch size of 12 and a learning rate set to 5×10^{-6}, utilizing 8 NVIDIA H100 GPUs. The AdamW optimizer (Loshchilov and Hutter 2018) is used with cross-entropy loss, and the maximum sequence length is set to 8192. Additionally, we derive an extra 64,000 QA training examples from the original training set, specifically tailored for content reiteration. For the Page Finder module, we utilize the `multi-qa-mpnet-base-dot-v1` model (Reimers and Gurevych 2021), and we further fine-tune it on 100,000 random samples generated from the training set. The fine-tuning process involves training for one epoch on 8 NVIDIA H100 GPUs, with a local batch size of 16, a learning rate set to 2×10^{-5}, and a warmup ratio of 0.1.

Evaluation Metrics We assess the generated answers using sentence-level automatic evaluation metrics provided by the Hugging Face evaluation pipeline (HuggingFace 2024). Automated metrics include BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE (Lin 2004), and BERTScore (Zhang et al. 2019), which provide quantitative measures of answer quality by comparing generated responses to reference answers. We also introduce the *page generation rate* and *page accuracy* to further demonstrate the effectiveness of our proposed page-finding optimization method. The *page generation rate* denotes the proportion of generated answers that include a reference to a page. The *page accuracy* refers to the proportion of these generated page references that correctly correspond to the intended reference pages.

Experimental Analysis

Impact of Context Window Size on Model Performance In our experiments with long documents, we investigate the effect of various context window sizes on model performance. Specifically, we use context window sizes of 0, 1, 3, and the maximum allowable size. Maximum window size includes as many adjacent pages as possible until reaching the maximum sequence length permitted by our experimental setup.

Results in Table 2 highlight the critical importance of context window size. Training without any context window (window size of 0) yields poorest performance, as conditioning on inputs without relevant context leads to unstable loss optimization. As context window size increases, the model's performance improves significantly, particularly in terms of *page accuracy* and *page generation rate*, as shown in Table 3. Incorporating a larger context window enhances the model's ability to accurately identify and generate relevant pages. These findings highlight the importance of using adjacent page information to improve model performance on multi-page document tasks, while balancing hardware constraints from memory and bandwidth for optimal results.

Enhancing Model Performance with Content Reiteration To further strengthen the model's page-finding capabilities, we implement a content reiteration strategy. This involves randomly sampling training examples and reformulating the QA pairs into a page-finding format. These reformulated samples are then incorporated into the original training set, enabling optimized instruction tuning. The experimental results in Table 2 indicate that content reiteration-guided instruction

Window Size	Content Reiteration	Page Finder	BLEU -1	BLEU -2	BLEU -3	BLEU -4	METEOR	ROUGE -1	ROUGE -2	ROUGE -L	ROUGE Lsum	BERT Score
/	×	×	65.9%	52.4%	45.2%	38.5%	60.9%	56.8%	42.2%	53.6%	53.7%	88.5%
/	×	✓	73.5%	63.7%	57.2%	48.6%	67.9%	63.9%	51.9%	61.7%	61.7%	90.7%
1	×	×	74.6%	62.8%	55.9%	49.7%	69.5%	66.1%	53.5%	63.7%	63.7%	91.7%
1	×	✓	73.6%	61.5%	54.4%	48.1%	67.8%	64.1%	51.1%	61.7%	61.7%	91.3%
1	✓	×	75.6%	64.1%	57.1%	51.0%	69.4%	66.5%	54.1%	64.3%	64.3%	91.8%
1	✓	✓	77.0%	66.3%	59.6%	53.7%	71.7%	68.9%	56.8%	66.5%	66.5%	92.3%
3	×	×	71.6%	61.1%	54.8%	48.7%	67.7%	63.5%	51.9%	61.2%	61.3%	90.6%
3	×	✓	73.3%	63.4%	57.3%	52.0%	70.5%	66.9%	55.5%	64.7%	64.7%	91.6%
3	✓	×	72.6%	62.0%	55.4%	49.5%	68.9%	64.9%	52.9%	62.6%	62.6%	91.1%
3	✓	✓	75.9%	65.9%	59.4%	53.6%	71.4%	68.1%	56.4%	65.8%	65.8%	91.8%
Max	×	×	72.3%	61.8%	55.3%	49.8%	70.6%	66.2%	54.1%	63.8%	63.8%	91.5%
Max	×	✓	73.6%	63.8%	57.6%	52.2%	72.4%	68.0%	56.4%	65.6%	65.6%	91.9%
Max	✓	×	75.1%	64.1%	57.4%	51.4%	69.7%	66.7%	54.4%	64.4%	64.4%	91.7%
Max	✓	✓	76.5%	66.2%	59.7%	54.0%	71.7%	68.7%	56.8%	66.4%	66.4%	92.2%

Table 2: Main results with different settings, including varying window sizes and the use of content reiteration or the *Page Finder* module. Window size refers to the number of adjacent pages included before and after the ground-truth page. The maximum window size denotes truncation centered around the ground-truth page, extending to the maximum sequence length allowed.

Window Size	Page Finding	Page Finder	Page Generation Rate	Page Accuracy
/	×	×	68.7%	83.2%
/	×	✓	88.2%	90.8%
1	×	×	69.7%	98.9%
1	×	✓	71.0%	99.1%
1	✓	×	73.9%	99.1%
1	✓	✓	83.4%	99.2%
3	×	×	80.4%	97.8%
3	×	✓	87.5%	98.5%
3	✓	×	79.2%	96.6%
3	✓	✓	88.1%	98.0%
Max	×	×	80.8%	98.8%
Max	×	✓	86.7%	99.2%
Max	✓	×	80.8%	97.6%
Max	✓	✓	88.5%	97.6%

Table 3: Page generation rate and page accuracy results. The generation rate represents the proportion of generated answers that include the reference page.

tuning significantly enhances the model's reading comprehension and page localization abilities. This approach proves particularly effective when the context window size is small, as presented in Table 3, by reinforcing the model's ability to pinpoint relevant pages. However, content reiteration alone may not fully compensate for the limitations of a small window size. Despite its benefits, the page accuracy for models trained with a small window size remains lower than that of models trained with the maximum window size, even without the use of content reiteration. This highlights that larger context windows are still crucial for optimal performance.

Impact of the Page Finder on Model Performance We further enhance our benchmark by incorporating the Page Finder module and conducting additional evaluations to verify its effectiveness, as shown in Table 2 and Table 3. The inclusion of the Page Finder allows us to mitigate practical limitations associated with processing very long documents, especially on mobile devices with constrained GPU memory. Inputting excessively long documents is often impractical due to hardware limitations and can introduce irrelevant information that may hinder the model's performance.

Moreover, the Page Finder facilitates better integration with retrieval-augmented generation (RAG) methods (Lewis et al. 2020; Izacard and Grave 2021). By providing grounding information and narrowing down the search space, it can aid subsequent research in developing more efficient and accurate models. Our design of the Page Finder within the QIDA-Quest offers a foundation for future studies to explore advanced techniques in document understanding and question answering tasks. Also, this groundwork paves the way for efficient deployment in real-world applications, enabling scalable solutions with rapid response.

Conclusion and Future Work

This paper introduces OIDA-QA, a multimodal QA benchmark designed specifically for OIDA document understanding. OIDA-QA comprises 400K training documents and 10K testing documents. In addition, we have collected over 370k multi-hop QA pairs generated by various models to enhance the dataset's diversity and robustness. Our experimental results demonstrate the effectiveness of both the benchmark and the AI assistant system in tackling multi-round QA tasks, presenting a promising approach to addressing opioid crisis.

Author Disclosure

Dr. Alexander is past Chair of FDA's Peripheral and Central Nervous System Advisory Committee and a co-founding Principal and equity holder in Stage Analytics. These arrangements have been reviewed and approved by Johns Hopkins University in accordance with its conflict-of-interest policies.

QA Generation Algorithm

Algorithm 1: QA Data Generation Process

Input : Document set $\mathcal{D}$;
Persona pool $\mathcal{P}$;
Desired number of QA pairs per document N_{QA};
Maximum attempts per document M

Output : QA dataset $\mathcal{QA}$

Initialize $\mathcal{QA} \leftarrow \emptyset$;
foreach *document* $D \in \mathcal{D}$ **do**
  Initialize QA count $n \leftarrow 0$, attempt count $m \leftarrow 0$;
  while $n < N_{QA}$ ***and*** $m < M$ **do**
    Sample personas $\mathcal{P}_D$ from $\mathcal{P}$ based on CLIP tags;
    foreach *persona* $P \in \mathcal{P}_D$ **do**
      Create prompt by combining document D and persona P;
      Generate question q using GPT-4o;
      if q *is answerable* **then**
        Generate answer a with referred page using GPT-4o;
        if (q, a) *pair is decomposable* **then**
          Decompose (q, a) into QA sequence $\{(q_i, a_i)\}_{i=1}^{s}$;
          if $s > 1$ **then**
            Add $\{(q_i, a_i)\}_{i=1}^{s}$ to $\mathcal{QA}$;
        $n \leftarrow n + 1$;
    $m \leftarrow m + 1$;

References

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Birnbaum, H. G.; White, A. G.; Schiller, M.; Waldman, T.; Cleveland, J. M.; and Roland, C. L. 2011. Societal costs of prescription opioid abuse, dependence, and misuse in the United States. *Pain medicine*, 12(4): 657–667.

CDC. 2020. *Americans Share Hopeful Stories of Recovery From Opioid Use Disorder*. https://www.cdc.gov/rxawareness/pdf/articles/TA-T3D2-English_MatteArticle_Release_508.pdf.

Chan, X.; Wang, X.; Yu, D.; Mi, H.; and Yu, D. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Da, C.; Luo, C.; Zheng, Q.; and Yao, C. 2023. Vision Grid Transformer for Document Layout Analysis. In *ICCV*.

Fox, S.; and Duggan, M. 2012. Health Online 2013.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gu, J.; Shi, X.; Kuen, J.; Qi, L.; Zhang, R.; Liu, A.; Nenkova, A.; and Sun, T. 2024. ADOPD: A Large-Scale Document Page Decomposition Dataset. In *ICLR*.

He, X.; Chen, S.; Ju, Z.; Dong, X.; Fang, H.; Wang, S.; Yang, Y.; Zeng, J.; Zhang, R.; Zhang, R.; et al. 2020. Meddialog: Two large-scale medical dialogue datasets. *arXiv preprint arXiv:2004.03329*.

Henderson, M.; Iyer, S.; Golub, D.; Agarwal, S.; Guo, D.; Huang, J.; Mayo, D.; Chen, R.; Tseng, B.; Siminian, M.; et al. 2017. Efficient Natural Language Response Suggestion for Smart Reply. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 331–339.

Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; and Wei, F. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. In *ACM Multimedia*.

HuggingFace. 2024. A library for easily evaluating machine learning models and datasets.

Izacard, G.; and Grave, E. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 874–880.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.; and Lu, X. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2567–2577.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Kuang, S.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 9459–9474.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81.

Liu, J.; Kong, Z.; Dong, P.; Shen, X.; Zhao, P.; Tang, H.; Yuan, G.; Niu, W.; Zhang, W.; Lin, X.; Huang, D.; and Wang, Y. 2025a. RoRA: Efficient Fine-Tuning of LLM with Reliability Optimization for Rank Adaptation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Liu, J.; Kong, Z.; Zhao, P.; Yang, C.; Shen, X.; Tang, H.; Yuan, G.; Niu, W.; Zhang, W.; Lin, X.; Huang, D.; and Wang, Y. 2025b. Toward Adaptive Large Language Models Structured Pruning via Hybrid-grained Weight Importance Assessment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18): 18879–18887.

Liu, Q.; Mao, R.; Geng, X.; and Cambria, E. 2023. Semantic matching in machine reading comprehension: An empirical study. *Information Processing & Management*, 60(2): 103145.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. *ICLR*.

Lu, Q.; Dou, D.; and Nguyen, T. 2022. ClinicalT5: A Generative Language Model for Clinical Text. In *EMNLP*.

NIDA. 2024. *Opioids*. https://www.drugabuse.gov/drug-topics/opioids.

Oderda, G. M.; Lake, J.; Rüdell, K.; Roland, C. L.; and Masters, E. T. 2015. Economic burden of prescription opioid misuse and abuse: a systematic review. *Journal of pain & palliative care pharmacotherapy*, 29(4): 388–400.

OIDA. 2021. Opioid Industry Documents Archive. https://www.industrydocuments.ucsf.edu/opioids/. Accessed: 2024-10-01.

OpenAI. 2022. Introducing chatgpt. https://openai.com/blog/chatgpt.

OpenAI, R. 2023. GPT-4 technical report. *arXiv*, 2303–08774.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, 248–260. PMLR.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Pergola, G.; Kochkina, E.; Gui, L.; Liakata, M.; and He, Y. 2021. Boosting Low-Resource Biomedical QA via Entity-Aware Masking Strategies. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1977–1985. Online: Association for Computational Linguistics.

Perot, V.; Kang, K.; Luisier, F.; Su, G.; Sun, X.; Boppana, R. S.; Wang, Z.; Mu, J.; Zhang, H.; and Hua, N. 2023. LMDX: Language Model-based Document Information Extraction and Localization. *arXiv preprint arXiv:2309.10952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982–3992.

Reimers, N.; and Gurevych, I. 2021. Sentence-Transformers: Multilingual Sentence, Paragraph, and Image Embeddings using BERT & Co. https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1. Accessed: 2025.

Shen, X.; Dong, P.; Lu, L.; Kong, Z.; Li, Z.; Lin, M.; Wu, C.; and Wang, Y. 2024a. Agile-Quant: Activation-Guided Quantization for Faster Inference of LLMs on the Edge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17): 18944–18951.

Shen, X.; Han, C.; Zhou, Y.; Xie, Y.; Gong, Y.; Wang, Q.; Wang, Y.; Wang, Y.; Zhao, P.; and Gu, J. 2025a. DraftAttention: Fast Video Diffusion via Low-Resolution Attention Guidance. *arXiv preprint arXiv:2505.14708*.

Shen, X.; Han, Z.; Lu, L.; Kong, Z.; Dong, P.; Li, Z.; Xie, Y.; Wu, C.; Leeser, M.; Zhao, P.; Lin, X.; and Wang, Y. 2024b. HotaQ: Hardware Oriented Token Adaptive Quantization for Large Language Models. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 1–1.

Shen, X.; Kong, Z.; Qin, M.; Dong, P.; Yuan, G.; Meng, X.; Tang, H.; Ma, X.; and Wang, Y. 2022. Data level lottery ticket hypothesis for vision transformers. *arXiv preprint arXiv:2211.01484*.

Shen, X.; Kong, Z.; Yang, C.; Han, Z.; Lu, L.; Dong, P.; Lyu, C.; Li, C.-h.; Guo, X.; Shu, Z.; et al. 2024c. Edgeqat: Entropy and distribution guided quantization-aware training for the acceleration of lightweight llms on the edge. *arXiv preprint arXiv:2402.10787*.

Shen, X.; Ma, W.; Liu, J.; Yang, C.; Ding, R.; Wang, Q.; Ding, H.; Niu, W.; Wang, Y.; Zhao, P.; Lin, J.; and Gu, J. 2025b. QuartDepth: Post-Training Quantization for Real-Time Depth Estimation on the Edge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11448–11460.

Shen, X.; Ma, W.; Zhou, Y.; Tang, E.; Xie, Y.; Li, Z.; Gong, Y.; Wang, Q.; Ding, H.; Wang, Y.; et al. 2025c. Fastcar: Cache attentive replay for fast auto-regressive video generation on the edge. *arXiv preprint arXiv:2505.14709*.

Shen, X.; Song, Z.; Zhou, Y.; Chen, B.; Li, Y.; Gong, Y.; Zhang, K.; Tan, H.; Kuen, J.; Ding, H.; Shu, Z.; Niu, W.; Zhao, P.; Wang, Y.; and Gu, J. 2025d. LazyDiT: Lazy Learning for the Acceleration of Diffusion Transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19): 20409–20417.

Shen, X.; Song, Z.; Zhou, Y.; Chen, B.; Liu, J.; Zhang, R.; Rossi, R. A.; Tan, H.; Yu, T.; Chen, X.; Zhou, Y.; Sun, T.; Zhao, P.; Wang, Y.; and Gu, J. 2025e. Numerical Pruning for Efficient Autoregressive Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19): 20418–20426.

Shen, X.; Wang, Y.; Lin, M.; Huang, Y.; Tang, H.; Sun, X.; and Wang, Y. 2023. DeepMAD: Mathematical Architecture Design for Deep Convolutional Neural Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6163–6173.

Shen, X.; Wang, Y.; Shi, X.; Wang, Y.; Zhao, P.; and Gu, J. 2025f. Efficient reasoning with hidden thinking. *arXiv preprint arXiv:2501.19201*.

Shen, X.; Zhao, P.; Gong, Y.; Kong, Z.; Zhan, Z.; Wu, Y.; Lin, M.; Wu, C.; Lin, X.; and Wang, Y. 2024d. Search for Efficient Large Language Models. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 139294–139315. Curran Associates, Inc.

Shen, X.; Zheng, H.; Gong, Y.; Kong, Z.; Yang, C.; Zhan, Z.; Wu, Y.; Lin, X.; Wang, Y.; Zhao, P.; and Niu, W. 2025g. Sparse Learning for State Space Models on Mobile. In *The Thirteenth International Conference on Learning Representations*.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.;

et al. 2022. Large Language Models Encode Clinical Knowledge. *arXiv preprint arXiv:2212.13138*.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; et al. 2023b. Towards Expert-Level Medical Question Answering with Large Language Models. *arXiv preprint arXiv:2305.09617*.

Swedo, E. A.; Sumner, S. A.; de Fijter, S.; Werhan, L.; Norris, K.; Beauregard, J. L.; Montgomery, M. P.; Rose, E. B.; Hillis, S. D.; and Massetti, G. M. 2020. Adolescent opioid misuse attributable to adverse childhood experiences. *The Journal of Pediatrics*, 224: 102–109.

Tanz, L. J.; Dinwiddie, A. T.; Mattson, C. L.; O'Donnell, J.; and Davis, N. L. 2022. Drug Overdose Deaths Among Persons Aged 10–19 Years - United States, July 2019-December 2021. *Morbidity and Mortality Weekly Report*.

Tu, Y.; Guo, Y.; Chen, H.; and Tang, J. 2023. LayoutMask: Enhance Text-Layout Interaction in Multi-modal Pre-training for Document Understanding. In *ACL*.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The dawn of lmms: Preliminary explorations with GPT-4V(ision). *arXiv preprint arXiv:2309.17421*, 9.

Yu, Y.; Li, Y.; Zhang, C.; Zhang, X.; Guo, Z.; Qin, X.; Yao, K.; Han, J.; Ding, E.; and Wang, J. 2023. StrucTexTv2: Masked Visual-Textual Prediction for Document Image Pre-training. In *ICLR*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Zhang, Y.; Zhang, R.; Gu, J.; Zhou, Y.; Lipka, N.; Yang, D.; and Sun, T. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.