LOCR: Location-Guided Transformer for Optical Character Recognition

Anonymous ACL submission

Abstract

Academic documents are packed with texts, equations, tables, and figures, requiring comprehensive understanding for accurate Optical Character Recognition (OCR). While end-to-end OCR methods offer improved accuracy over layout-based approaches, they often grapple with significant repetition issues, especially with complex layouts in Out-Of-Domain (OOD) documents. To tackle this issue, we propose LOCR[1], a model that integrates location guiding into the transformer architecture during autoregression. We train the model on an original large-scale dataset comprising over 53M text-location pairs from 89K academic document pages, including bounding boxes for words, tables and mathematical symbols. LOCR adeptly handles various formatting elements and generates content in Markdown language. It outperforms all existing methods in our test set constructed from arXiv, as measured by edit distance, BLEU, METEOR and F-measure. LOCR also eliminates repetition in the arXiv dataset, and reduces repetition frequency in OOD documents, from 13.19% to 0.04% and from 8.10% to 0.11% for natural science and social science documents respectively. Additionally, LOCR features an interactive OCR mode, facilitating the generation of complex documents through a few location prompts from human.

1 Introduction

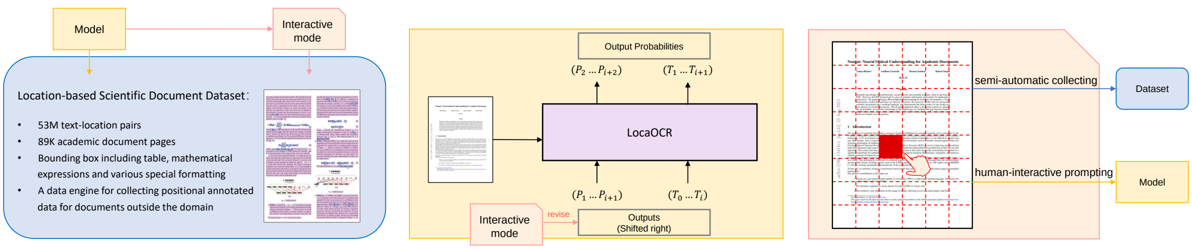
Academic literature comprises a wealth of high-quality content, yet much of it is provided in formats like PDF that are not readily for machine reading. Particularly, most academic documents of the previous centuries are scanned version. Digitizing academic documents are important for scientific research, literature retrieval, and large-language model training. However, academic document layout tends to be highly intricate, including text, equations, images, tables, and annotations, posing challenges for obtaining accurate OCR results.

One approach to document OCR is to first analyze the layout of the document and then extract the text content (Zhu et al., 2022,mindee, 2023). While progress has been made in any of the two stages or handling specific types of elements, such as table detection and recognition (Yang et al., 2022), handwritten formula recognition (Sakshi and Kukreja, 2023) and structured information extraction (Lu et al., 2022; Liao et al., 2023), it is very difficult for models to understand all the elements and connect the different chunks into a coherent sequence.

Recently, an end-to-end transformer structure, Donut (Kim et al., 2022), was proposed for document understanding. It effectively addresses the complexity of combining multiple models and the issue of error propagation. Without too many changes in the model, Nougat (Blecher et al., 2023) processes academic PDFs into markup language. However, these methods are prone to hallucination and repetitions, such as continuously repeating the same sentence on a page.

In fact, getting trapped in a repetitive loop is a common problem with Transformer-based models sampling with greedy search decoding (Holtzman et al., 2019). It is challenging for a language model to accurately capture all the content of text-intensive documents without position perception. By visualizing the cross-attention during the prediction process of Nougat (see Appendix E), we found that the cross-attention cannot be focused on the correct position when the layout is complex. This phenomenon indicates that the positional information influence the text decoding to a great extent. Inspired by this, we consider incorporating positional guidance for the model to focus on the correct word to address the issue of repetitive loop. We introduce

[1]Source codes and datasets will be available under the MIT license upon publication



(a) **Data**: dataset & data engine (b) **Model**: location-guided transformer (c) **Interactive**: align with human intent

Figure 1: An overview of three components of our work: a ***large-scale dataset*** with positional annotation and a data engine, a ***location-guided OCR model*** for various layouts, and an ***interactive mode*** for humans to prompt the model and modify data collection.

LOCR, a location-guided document understanding model, together with an original large-scale dataset and an interactive OCR mode to align with human intention (see Figure 1 for an overview).

The most significant feature that distinguishes our model from previous works is the incorporation of positional autoregression alongside text autoregression. LOCR simultaneously predicts the current token and the position of the next token, which is used to prompt the decoding of the next token. Through this method, we not only combine positional information with text information but also avoid the tedious process and error accumulation in the two-stage OCR method. Taking document images as input, our model outputs document content in Markdown format, including special formats such as superscripts and subscripts.

Furthermore, we propose an importance decay strategy to intuitively penalize locations that have already been visited to avoid repetition. With the record of visited locations, we decrease the importance of these positions. The repetition behavior is eliminated in the arXiv test set, and decreases for out-of-domain documents. For documents with complex layouts, we also introduce an interactive OCR mode, allowing the model to continue to decode the text where the user has dragged a box. With these enhancement strategies, the generation ability of the model is significantly improved.

Additionally, we propose a data engine for constructing academic document OCR dataset with positional annotations. We collect a large-scale dataset of 89K academic document pages with 53M text-location pairs. To the best of our knowledge, it is the first dataset that includes a bounding box of each mathematical symbol in academic documents.

In summary, the contributions of this paper are:

- We introduce LOCR, a transformer-structured OCR model with positional supervision. Our model achieves the state-of-the-art score in academic document understanding task in the arXiv test set (see Section 5.2) and alleviates the repetitive degradation to a great extent (Section 5.3).
- We innovatively introduce an interactive OCR mode, enabling the model to handle any out-of-domain documents. Humans only need to provide the position box for the next word without any cumbersome operations (see Section 5.5).
- We will release a large-scale dataset composed of 89K pages of academic documents. Each piece of data contains a document page image, the texts in Markdown format, and the bounding boxes of all words and mathematical symbols (see Section 3).

2 Related Work

2.1 General-purpose OCR

Optical Character Recognition (OCR) caters to a diverse array of applications, including document digitization (Smith, 2007; Moysset et al., 2017), handwriting recognition, and scene text recognition (Li et al., 2021; Bautista and Atienza, 2022). The classic OCR methods consist of two stages: text detection and text recognition. The text detection algorithm obtains the position of text boxes from the image, and then the recognition algorithm recognizes the content within the text boxes. Researches in these sub-fields have achieved satisfactory results, such as EAST (Zhou et al., 2017) for text detection, CRNN (Shi et al., 2015) for text recognition, and LayoutLM family (Xu et al., 2019; Xu et al., 2020; Huang et al., 2022) for document element identification. There also has been various integrated toolbox to connect the above functions, such as DocXChain (Yao, 2023) and EffOCR (Bryan et al., 2023).

2.2 Academic document OCR

For academic document understanding, additional tasks like table and mathematical equation parsing are also involved. Marker (Paruchuri and Lampa, 2023) is a pipeline of text extracting, layout detection, and block combination, which converts PDF, EPUB, and MOBI to Markdown with a series of deep learning models. PaddleOCR develops a document analysis system PP-Structure (Li et al., 2022), which first analyses the layout information and then extracts key information. Such OCR-based approaches have shown promising performance but suffer from complexity and error propagation to the subsequent process. To address this issue, document understanding models based on transformer structure were proposed. Donut (Kim et al., 2022) is an encoder-decoder model that directly decodes the expected sequences from visual inputs. Nougat (Blecher et al., 2023) is a specific model trained on academic documents to process academic PDFs into markup language, with the ability to parse images of math equations and tables.

With the emergence of general large models, some Large Vision-Language Models (LVLMs) mark a significant milestone across OCR tasks. Vary (Wei et al., 2023) is a document parsing method, equipping the large model with the fine-grained perception and understanding by scaling up the vision vocabulary of LVLMs. As the state-of-the-art LVLM, GPT-4v (Yang et al., 2023) performs well in recognizing and understanding Latin contents. But it shows limitations when dealing with complex tasks such as table structure recognition and semantic entity recognition (Shi et al., 2023). When it comes to unstructured layouts or inconsistent text distribution, GPT-4v tends to omit lengthy tables and only reconstruct the short beginning of that.

Without the box detection of two-stage OCR, the methods above are prone to hallucination and repetitions. This phenomenon indicates that it is crucial for the model to find the correct position in order to generate the correct sequences, especially for ambiguous layouts and out-of-domain documents.

2.3 Promptable model

Interactive models play a significant role in aligning behavior of artifical intelligence with human intentions, which have shown promising performance within a variety of domains. SAM(Kirillov et al., 2023) presents an interactive segmentation model capable of accommodating point, box, and text-based input. DINOv (Li et al., 2023) achieves visual in-context prompting in both referring and general segmentation. T-Rex (Jiang et al., 2023) explores object detection and counting, which can interactively refine the counting results by prompting on missing or falsely-detected objects. In contrast, the field of OCR revolves less interactive explorations, despite the dealing with complex layout has an urge for human prompts and interactions.

3 Dataset

3.1 Data collection

To the best of our knowledge, there is no paired dataset containing markup-formatted document contents along with corresponding bounding boxes (bbox) for each word and mathematical symbol. We proposed a data engine to collect such paired data. The process is shown in Figure 2.

We get the Tex source files of academic papers from arXiv. In the first step, we assign a unique RGB color identifier to each word and mathematical symbol automatically by using xcolor package in LaTeX (see Step1). In the second step, following the same pipeline as Nougat (Blecher et al., 2023), we compile LaTeX files into PDF and Markdown files respectively. Since PDF is a rich text format that supports color changes, we obtain colorful PDF files. While Markdown is a plain text format, the RGB identifiers are compiled into text forms (see Step2). In the third step, we use the PyMuPDF package of python to parse the colorful PDF files and extract the pair of (color, bbox). At the same time, we parse the Markdown file with regular expressions to get the paired (color, text) data. Finally, we merge the two pairs of data by the key of RGB color to get paired (text, bbox) data (see Step3).

We collected academic papers released on arXiv from 2007 to 2023. During data processing, some articles failed the conversion due to user-defined macros or non-standardized formats. After all conversion and data cleaning, our dataset is composed of 88998 pages, which include, but are not limited to, the bounding box of plain text, Greek letters, arithmetic symbols, superscripts, subscripts, and tabular symbols. Examples of our dataset is avail-

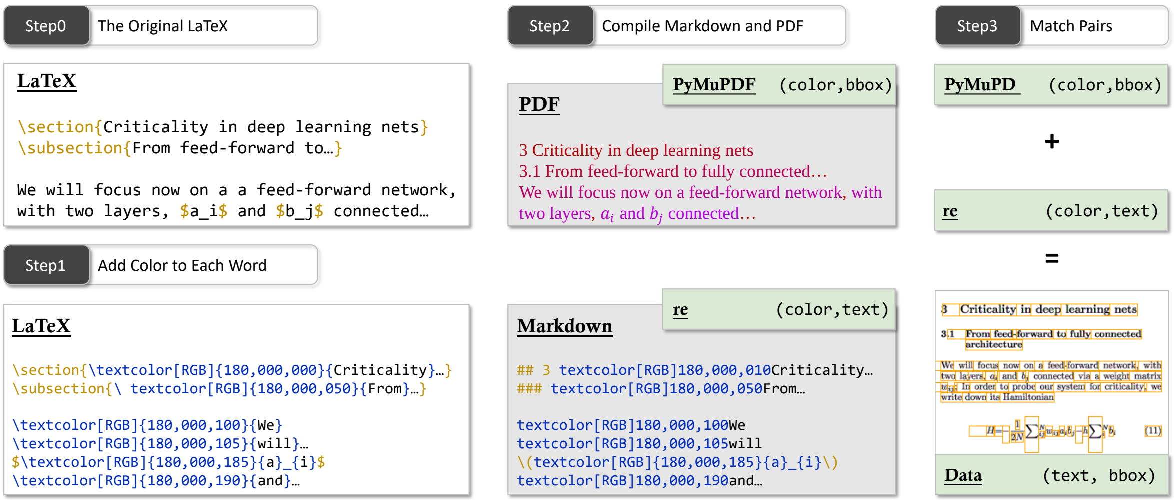


Figure 2: Data Processing. Step1: Add a unique RGB identifier to each word by parsing the Tex file. Step2: Convert source file into Markdown and PDF formats respectively. Step3: Extract color-bbox pairs from colored PDF, color-text pairs from Markdown, and merge the two to get the text-bbox pairs.

able in Appedix A1.

3.2 Data augmentation

Image augmentation To simulate the imperfections and variability of scanned documents, we follow (Simard et al., 2003) to apply data augmentation to document images, including of erosion, dilation, gaussian noise, gaussian blur, bitmap conversion, image compression, grid distortion and elastic transform. Each of the transformations is applied with a certain probability.

Text augmentation To address the issue of the model getting stuck in repetitive loops, we randomly skip 0 to 5 tokens and their corresponding positions in the ground truth labels. Compared with the perturbation method in Nougat, which randomly replaces tokens, our method shows a more pronounced effect (see Section 5.3).

Position augmentation Since bounding boxes are involved in the autoregressive process, there may be some imprecise output. In some cases, a user may also draw a loose box in the interactive mode. Therefore, it is reasonable to add noise to the bounding boxes during the training phase. We add Gaussian noise with a mean of 0 and a standard deviation of 0.5 times the side length to each box.

4 Methodology

4.1 Model structure

The over view of our model is shown in Figure 3, with a transformer-based backbone and an additional prompt module to process positional information. Given an image as input, the image encoder transforms it as image embedding. Semantic information and visual information are integrated within the decoder, enabling simultaneous prediction of the current token and its next position.

Backbone Theoretically, our prompt module can be applied to any multimodal models with an image encoder and a text decoder. When no positional information is provided, the backbone model would autonomously generate sequences. In this paper, we choose Nougat (Blecher et al., 2023) as the backbone, which uses the implementation of Swin Transformer (Liu et al., 2021) as image encoder and mBART (Lewis et al., 2019) as decoder. Given an image of $x \in R^{3,H_0,W_0}$, the image encoder transfers it into dense embedding $h_{img} \in R^{H,W,d}$, which is then decoded into a sequence of token embeddings $h_t \in R^d$. Finally, the sequence of token embeddings is projected into a logit matrix with the size of the vocabulary v.

Prompt Module Without location guiding, the backbone model may get confused about where to find the next token. The prompt module is designed to perceive spatial information prompted by previous steps or human, consisting of two-dimensional positional encoding and position detection heads.

We opt for positional encodings with Fourier Features (Tancik et al., 2020) to represent the positions of bounding boxes for both tokens and the image. The token bounding box, defined by its top-left and

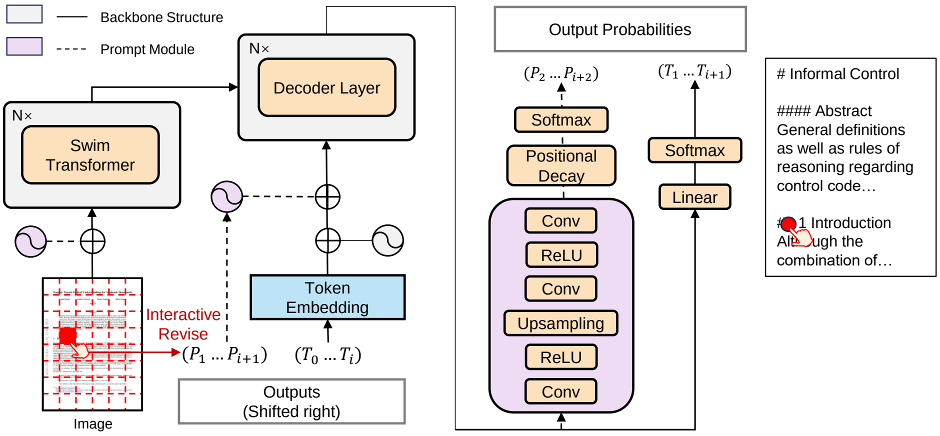


Figure 3: Model Architecture. Left: Image encoder and decoder of transformer structure. Right: Position detection head and token projection. Purple: Prompt module consisting of positional encodings and position detection head. Red: Interactive mode with human-reviewed input.

bottom-right corners, is transformed into a dense position embedding $h_{box} \in R^d$. For the image embedding $h_{img} \in R^{H,W,d}$, we divide it into grids of size (H, W) (shown in Figure 3), and apply positional encodings to each grid box to get the its position embedding $h_{grid} \in R^{H,W,d}$.

The position detection heads are used to predict the position of the next token. Given that the weights of the cross-attention layers indicate the similarity between image grids and the current token, we utilize them as input for position detection. Inspired by CenterNet (Duan et al., 2019), an effective object detection algorithm, we use three convolutional heads to predict the position of the next token. The first convolution head predicts the grid containing the next token by conducting a classification task on all grids in an image. The second and third convolution heads regress the size and center offset of the next bounding box respectively. Finally, the coordinates of the bounding box are calculated based on the center point and the width and height. To improve prediction accuracy, we upsample the image grid output by decoder from (H,W) to (2H,2W), allowing finer-grained positition prediction.

Information fusion The token information and spatial information is fused in cross-attention layers of decoder. In backbone models without prompt module, the cross-attention layers take solely image embedding as encoder hidden states and token embedding as hidden states input. Instead, we use the sum of the image embedding $h_{img} \in R^{H,W,d}$ and its position embedding $h_{grid} \in R^{H,W,d}$ as the encoder hidden states, and the sum of token embedding $h_t \in R^d$ and position embedding $h_{box} \in R^d$ as the hidden states input. As a consequence, in cross-attention layers where token information interacts with the image contents, the positional information of tokens and image are also fused.

4.2 Decay strategy for anti-repetition

During the inference stage, we introduce position decay strategies based on prior knowledge to guide the prediction of positions.

Accumulation Decay The accumulation decay strategy is implemented by recording the count of tokens that have appeared in each grid. The heatmap for predicting the next grid is adjusted by penalizing grids where many tokens have already been located as follows:

$$hm = hm + log(\sigma) \cdot cnt \qquad (1)$$

Where $hm \in R^{2H,2W}$ denotes the upsampled heatmap predicted by the first position detection head and $cnt \in R^{2H,2W}$ denotes the count of tokens that have appeared in each grid. The $\sigma \in (0, 1]$ denotes decay rate. Smaller σ value means stronger decay effect. When σ is set to 1, the decay function is deactivated. We recommend using a decay rate between 0.75 and 0.95, depending on the density of text in the target documents.

Blank Decay Another intuitive idea is to apply positional decay to blank grids. We calculate the standard deviation std for pixels within each grid, where grids with smaller standard deviations (in extreme cases, containing no characters at all) are considered less likely to contain the next token.

Together with blank decay strategy, the heatmap is adjusted as follows:

$$hm = hm + log(\sigma) \cdot cnt + log(\eta \cdot std) \quad (2)$$

4.3 Loss function

Our loss function consists of two parts: token loss and position loss.

Token loss We use the cross-entropy loss of tokens L_t to train the language decoder.

Position loss For the three convolutional heads in the position detection module, we apply cross-entropy loss to the first classification head and the Intersection over Union (IOU) metric to the subsequent two heads. Additionally, we integrate the normalized Euclidean distance between the center of the predicted box and that of the target box to mitigate the shortcomings of slow convergence and inaccurate regression inherent in IOU (Zheng et al., 2019). The position loss function is as follows:

$$L_p = \alpha L_p^{ce} + \beta(1 - iou + \gamma d^2) \quad (3)$$

Where L_p^{ce} denotes the cross-entropy loss of the classification. d represents the normalized Euclidean distance to adjust the IOU loss. Additionally, α, β, and γ are hyperparameters, corresponding to 1, 0.3, and 10 respectively in our settings.

As the prediction of the text at the beginning of a page is much more challenging and important, we assigned a higher weight θ for the initial text than the subsequent text.

The final loss function is as follows:

$$l = \theta(L_p^{init} + L_t^{init}) + L_p^{sub} + L_t^{sub} \quad (4)$$

4.4 Human interaction

As a complement to our method, we provide an interactive mode, which serves both for improving the model's performance and as a part of our data construction engine.

Model Assistant To deal with extremely hard cases, we provide a browser-based tool to enable users to give real-time position prompts by simply dragging a box. When the autoregressive process encounters a state of confusion, characterized by a predicted token or position confidence lower than a predetermined threshold, users can opt to provide a positional prompt. With the correct position provided, the autoregressive process would go on more smoothly (see Section 5.5 for results).

Data construction With the model automatically predicting positions, minimal human intervention is required to acquire additional out-of-domain data, particularly the positional bounding box labels. As a result, LOCR is able to parse a broader range of layouts and document domains beyond academic papers. For instance, when tested on patent documents, LOCR's recognition of the majority of content is satisfactory (see Figure B4), showing the model's flexibility. This paves the way for broader applications of location-based OCR method.

5 Result and Evaluation

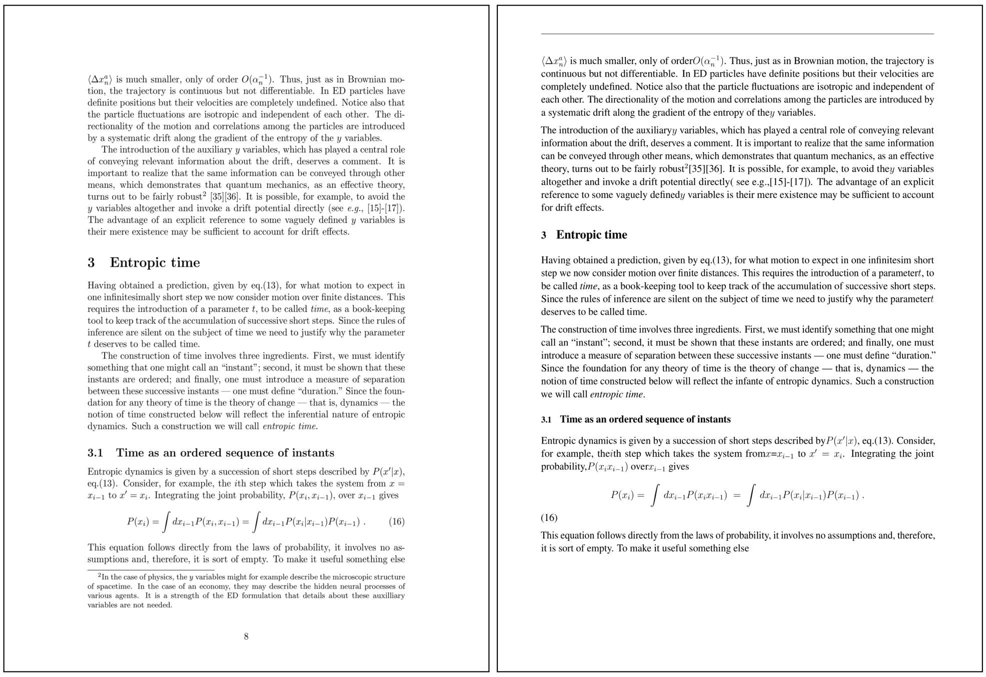
5.1 Implementation details

Baseline We use both the state-of-the-art integrated toolbox Marker, PaddleOCR and end-to-end generation model Nougat as our baselines. For PaddleOCR, which outputs each bounding box by text detection and corresponding text by text recognition, we concatenate the sequences in the order of its model output.

Dataset Since our main baseline model, Nougat, does not provide an open resource dataset, we evaluate our method with the dataset introduced in Section 3, which shares the same data source and processing pipeline as Nougat. The test set contains 1000 pages of academic documents. In the testing phase, only images are used as inputs, which ensures the fairness and rationality of our evaluation.

Setup We resize the input dimensions of the images to (H_0, W_0) = (896, 672), an aspect ratio that accommodates the majority of academic paper sizes. The maximal sequence length of transformer decoder is set to 4096 to allow the output of intensive text in academic research papers. During inference the text is generated using greedy decoding.

Training details We initialize the backbone parameters using the pretrained Nougat small model, while the prompt module is initialized randomly. LOCR was trained for 50 epochs using 64 A100 80GB GPUs, with a total batch size of 128. The maximum learning rate is set to 5×10^{-4}, with exponential decay until reaching 1×10^{-5}.



$\langle \Delta x_n^a \rangle$ is much smaller, only of order $O(\alpha_n^{-1})$. Thus, just as in Brownian motion, the trajectory is continuous but not differentiable. In ED particles have definite positions but their velocities are completely undefined. Notice also that the particle fluctuations are isotropic and independent of each other. The directionality of the motion and correlations among the particles are introduced by a systematic drift along the gradient of the entropy of the y variables.

The introduction of the auxiliary y variables, which has played a central role of conveying relevant information about the drift, deserves a comment. It is important to realize that the same information can be conveyed through other means, which demonstrates that quantum mechanics, as an effective theory, turns out to be fairly robust[2] [35][36]. It is possible, for example, to avoid the y variables altogether and invoke a drift potential directly (see *e.g.*, [15]-[17]). The advantage of an explicit reference to some vaguely defined y variables is their mere existence may be sufficient to account for drift effects.

3 Entropic time

Having obtained a prediction, given by eq.(13), for what motion to expect in one infinitesimally short step we now consider motion over finite distances. This requires the introduction of a parameter t, to be called *time*, as a book-keeping tool to keep track of the accumulation of successive short steps. Since the rules of inference are silent on the subject of time we need to justify why the parameter t deserves to be called time.

The construction of time involves three ingredients. First, we must identify something that one might call an "instant"; second, it must be shown that these instants are ordered; and finally, one must introduce a measure of separation between these successive instants — one must define "duration." Since the foundation for any theory of time is the theory of change — that is, dynamics — the notion of time constructed below will reflect the inferential nature of entropic dynamics. Such a construction we will call *entropic time*.

3.1 Time as an ordered sequence of instants

Entropic dynamics is given by a succession of short steps described by $P(x'|x)$, eq.(13). Consider, for example, the ith step which takes the system from $x = x_{i-1}$ to $x' = x_i$. Integrating the joint probability, $P(x_i, x_{i-1})$, over x_{i-1} gives

$$P(x_i) = \int dx_{i-1} P(x_i, x_{i-1}) = \int dx_{i-1} P(x_i|x_{i-1}) P(x_{i-1}) \ . \qquad (16)$$

This equation follows directly from the laws of probability, it involves no assumptions and, therefore, it is sort of empty. To make it useful something else

[2]In the case of physics, the y variables might for example describe the microscopic structure of spacetime. In the case of an economy, they may describe the hidden neural processes of various agents. It is a strength of the ED formulation that details about these auxilliary variables are not needed.

8

$\langle \Delta x_n^a \rangle$ is much smaller, only of order$O(\alpha_n^{-1})$. Thus, just as in Brownian motion, the trajectory is continuous but not differentiable. In ED particles have definite positions but their velocities are completely undefined. Notice also that the particle fluctuations are isotropic and independent of each other. The directionality of the motion and correlations among the particles are introduced by a systematic drift along the gradient of the entropy of the$y$ variables.

The introduction of the auxiliary$y$ variables, which has played a central role of conveying relevant information about the drift, deserves a comment. It is important to realize that the same information can be conveyed through other means, which demonstrates that quantum mechanics, as an effective theory, turns out to be fairly robust[2][35][36]. It is possible, for example, to avoid the$y$ variables altogether and invoke a drift potential directly(see e.g.,[15]-[17]). The advantage of an explicit reference to some vaguely defined$y$ variables is their mere existence may be sufficient to account for drift effects.

3 Entropic time

Having obtained a prediction, given by eq.(13), for what motion to expect in one infinitesim short step we now consider motion over finite distances. This requires the introduction of a parameter$t$, to be called *time*, as a book-keeping tool to keep track of the accumulation of successive short steps. Since the rules of inference are silent on the subject of time we need to justify why the parameter$t$ deserves to be called time.

The construction of time involves three ingredients. First, we must identify something that one might call an "instant"; second, it must be shown that these instants are ordered; and finally, one must introduce a measure of separation between these successive instants — one must define "duration." Since the foundation for any theory of time is the theory of change — that is, dynamics — the notion of time constructed below will reflect the infante of entropic dynamics. Such a construction we will call *entropic time*.

3.1 Time as an ordered sequence of instants

Entropic dynamics is given by a succession of short steps described by$P(x'|x)$, eq.(13). Consider, for example, the$i$th step which takes the system from$x$=x_{i-1} to $x' = x_i$. Integrating the joint probability,$P(x_i x_{i-1})$ overx_{i-1} gives

$$P(x_i) = \int dx_{i-1} P(x_i x_{i-1}) = \int dx_{i-1} P(x_i|x_{i-1}) P(x_{i-1}) \ .$$

(16)

This equation follows directly from the laws of probability, it involves no assumptions and, therefore, it is sort of empty. To make it useful something else

Figure 4: Examples of our model output. Left: Origin image of document page. Right: Model output converted to Markdown and rendered back into a PDF. More detailed examples are available in Appendix B

Method	Edit dist↓	BLEU↑	METEOR↑	Precision↑	Recall↑	F1↑
PaddleOCR	0.475	0.500	0.589	0.713	0.690	0.696
Marker	0.221	0.696	0.783	0.838	0.804	0.814
Nougat small (247M*)	0.166	0.825	0.882	0.900	0.898	0.899
Nougat base (348M*)	0.159	0.829	0.889	0.900	0.905	0.902
LOCR (248M*,$\sigma = 1$)	0.106	**0.854**	**0.913**	**0.915**	**0.916**	**0.915**
LOCR (248M*,$\sigma = 0.85$)	**0.104**	**0.854**	0.912	**0.915**	0.915	**0.915**
LOCR (248M*,$\sigma = 0.75$)	0.109	0.850	0.910	0.914	0.911	0.912

Table 1: Comparative performance results on the arXiv test set. Our LOCR method demonstrates superior performance across multiple metrics, significantly outperforming the baseline methods. *Number of parameters.

5.2 Metrics

Following Nougat (Blecher et al., 2023), we use Edit distance, BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), Precision, Recall and F1 to measure the quality of output text.

As shown in Table 1, while the number of LOCR's parameters is only slightly more than the small version of Nougat, our model outperforms the base version of Nougat in all evaluation metrics. In contrast, the multi-stage pipelines do not convert all equations to LaTeX and not all lines are joined properly. For the autogressive method without position supervision, Nougat prones to hallucination and repetition. These results confirm the effectiveness of LOCR and the positional decay strategy.

Besides, we use IOU metrics to measure the performance of our prompt module. LOCR achieves a IOU score of 0.702. Our method successfully handles various layouts, including pages with multiple subfigures, tables, mathematical formulas, and references (Examples are available in Appendix B).

5.3 Repetition

Following Nougat (Blecher et al., 2023), we detect the repetition behavior during inference by computing the variances of the largest logit values of each step. If the signal drops below a threshold, we regard the sequence to have repetitions.

We evaluate the generation ability of our model and present the frequency of repetition in Table 2. Due to the majority of arXiv manuscripts being formatted in single or double columns and lack-

Method	ArXiv			Quantum			Marketing		
	Page	Doc*	Cover	Page	Doc*	Cover	Page	Doc*	Cover
Nougat small	4.39%	27.60%	6.40%	13.77%	63.90%	22.70%	8.30%	60.80%	14.50%
Nougat base	4.42%	27.80%	5.30%	13.19%	55.40%	15.40%	8.10%	60.20%	16.90%
LOCR ($\sigma = 1$)	0.88%	5.20%	0.30%	2.78%	17.10%	0.60%	1.36%	11.90%	0.70%
LOCR ($\sigma = 0.85$)	0.01%	0.10%	0.10%	0.08%	0.60%	**0.00%**	**0.11%**	**1.40%**	**0.00%**
LOCR ($\sigma = 0.75$)	**0.00%**	**0.00%**	**0.00%**	**0.04%**	**0.30%**	**0.00%**	0.14%	1.60%	0.10%

Table 2: Robustness of LOCR across diverse domains, showcasing the significant reduction in generation failures. The three columns for each domain are calculated based on failed pages / total pages, failed doc / total doc, and doc with failed cover / total doc. *Statistics on the number of pages in each document can be found in Appendix D.

ing complex layout such as footnotes and covers, we selected out-of-domain (OOD) datasets from diverse fields to ensure varied layouts. Specifically, we select 1000 papers each from natural sciences (quantum physics) and social sciences (marketing), as OOD test documents. We calculate both the proportion of failed pages and that of failed documents. As the first page of an academic document typically shows a more complex layout than the subsequent pages, we additionally calculate the proportion of documents with failures in the cover.

The model exhibits an impressive decrease in repetition failures. Specifically, in arXiv dataset, LOCR with $\sigma = 0.75$ eliminates repetition for all pages from 4.42%. For OOD documents where the documents are more challenging to comprehend with more complex formulas, LOCR with $\sigma = 0.75$ reduces the failure rate for all pages to 0.04% for quantum documents and LOCR with $\sigma = 0.85$ reduces that to 0.11% for marketing documents. On the other hand, among all failed documents, the proportion of failures on the first page is significantly decreased, demonstrating better ability of LOCR to handle more complex layouts. Some pages that failed with Nougat but were successfully converted by LOCR are shown in Appendix B.

5.4 Ablation study

We conduct ablation study to illustrate the individual contribution of the decay strategy and the positional module.

Regarding the decay strategy, the bottom three rows in Table 1 preliminarily demonstrate its efficacy, where $\sigma = 1$ signifies no decay strategy applied. Further, we conducted ablation experiments on the repetition rate. As Table 2 shows, our decay strategy proves further performance improvement compared to scenarios without the decay strategy. Besides, the model results show good robustness to slight fluctuations of decay rate.

Regarding the positional module, comparing the performance of LOCR with that of the Nougat model serves as a valuable ablation experiment. Since our final training set constitutes a subset of Nougat's training set, in the absence of the decay strategy ($\sigma = 1$) in Table 1, the performance improvement of our model serves as evidence of the effectiveness of the positional module.

5.5 Interaction

Although the problem of repetitive degeneration has been largely alleviated, we aim to complete the remaining layouts in the interactive mode. When the model encounters a layout that is difficult to judge and the confidence of the predicted position is lower than the threshold, simply dragging a bounding box allows the model to automatically return to the expected position and continue outputting correct results. Interactive examples are available in Appendix C. As a result, LOCR is able to parse a broader range of document domains beyond academic papers. An example of LOCR to parse patent documents is shown in Figure B4.

6 Discussion

In our work, we introduce LOCR, which incorporates location guiding into the language model. Our approach significantly mitigates the problem of repetitive loops often encountered by transformer-based models. The interactive mode can be utilized to construct datasets for fine-tuning OCR models to specific domain literature, and enhancing the generalization capability of our model. We believe that LOCR can be applied to digitize documents from various fields with complex layouts, thereby assisting academic research, literature retrieval, and large language model training. We hope this work can help the development of the area of OCR.

7 Limitations

Although the frequency of repetition has been significantly mitigated, it has not been entirely eradicated in out-of-domain documents. Secondly, when parsing other types of documents beyond academic papers, some human interaction is needed. Additionally, our model encounters difficulties when the initial word on a page is incomplete, leading to imperfect handling. We will continue our work to address these issues.

References

Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72.

Darwin Bautista and Rowel Atienza. 2022. Scene Text Recognition with Permuted Autoregressive Sequence Models. *arXiv e-prints*, arXiv:2207.06966.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

Tom Bryan, Jacob Carlson, Abhishek Arora, and Melissa Dell. 2023. EfficientOCR: An Extensible, Open-Source Package for Efficiently Digitizing World Knowledge. *arXiv e-prints*, arXiv:2310.10050.

Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. *arXiv e-prints*, arXiv:2204.08387.

Qing Jiang, Feng Li, Tianhe Ren, Shilong Liu, Zhaoyang Zeng, Kent Yu, and Lei Zhang. 2023. T-rex: Counting by visual prompting. *arXiv preprint arXiv:2311.13596*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chenxia Li, Ruoyu Guo, Jun Zhou, Mengtao An, Yuning Du, Lingfeng Zhu, Yi Liu, Xiaoguang Hu, and Dianhai Yu. 2022. PP-StructureV2: A Stronger Document Analysis System. *arXiv e-prints*, arXiv:2210.05391.

Feng Li, Qing Jiang, Hao Zhang, Tianhe Ren, Shilong Liu, Xueyan Zou, Huaizhe Xu, Hongyang Li, Chunyuan Li, Jianwei Yang, et al. 2023. Visual in-context prompting. *arXiv preprint arXiv:2311.13601*.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. *arXiv e-prints*, arXiv:2109.10282.

Haofu Liao, Aruni RoyChowdhury, Weijian Li, Ankan Bansal, Yuting Zhang, Zhuowen Tu, Ravi Kumar Satzoda, R Manmatha, and Vijay Mahadevan. 2023. Doctr: Document transformer for structured information extraction in documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19584–19594.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. *arXiv preprint arXiv:2203.12277*.

mindee. 2023. doctr: Document text recognition. `https://github.com/mindee/doctr`.

Bastien Moysset, Christopher Kermorvant, and Christian Wolf. 2017. Full-Page Text Recognition: Learning Where to Start and When to Stop. *arXiv e-prints*, arXiv:1704.08628.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Vik Paruchuri and Samuel Lampa. 2023. Marker: Convert pdf to markdown quickly with high accuracy. `https://github.com/VikParuchuri/marker?tab=readme-ov-file`.

Sakshi Sakshi and Vinay Kukreja. 2023. Recent trends in mathematical expressions recognition: An lda-based analysis. *Expert Systems with Applications*, 213:119028.

Baoguang Shi, Xiang Bai, and Cong Yao. 2015. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition. *arXiv e-prints*, arXiv:1507.05717.

Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. 2023. Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation. *arXiv preprint arXiv:2310.16809*.

P.Y. Simard, D. Steinkraus, and J.C. Platt. 2003. Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 958–963.

R. Smith. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. 2020. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. *arXiv e-prints*, arXiv:2006.10739.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2023. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2020. LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding. *arXiv e-prints*, arXiv:2012.14740.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2019. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. *arXiv e-prints*, arXiv:1912.13318.

Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel, and Shachi Paul. 2022. Tableformer: Robust transformer modeling for table-text encoding. *arXiv preprint arXiv:2203.00274*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv e-prints*, arXiv:2309.17421.

Cong Yao. 2023. DocXChain: A Powerful Open-Source Toolchain for Document Parsing and Beyond. *arXiv e-prints*, arXiv:2310.12430.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. 2019. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *arXiv e-prints*, arXiv:1911.08287.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. EAST: An Efficient and Accurate Scene Text Detector. *arXiv e-prints*, arXiv:1704.03155.

Wenzhen Zhu, Negin Sokhandan, Guang Yang, Sujitha Martin, and Suchitra Sathyanarayana. 2022. DocBed: A Multi-Stage OCR Solution for Documents with Complex Layouts. *arXiv e-prints*, arXiv:2202.01414.

A Dataset Examples

To the best of our knowledge, this is the first paired dataset containing markup-formatted document contents along with corresponding bounding boxes. What makes our dataset distinguished from existing ones is that our bounding boxes covers all visible mathematical symbols, such as $\sum$, $\langle\rangle$ and θ^{α}.

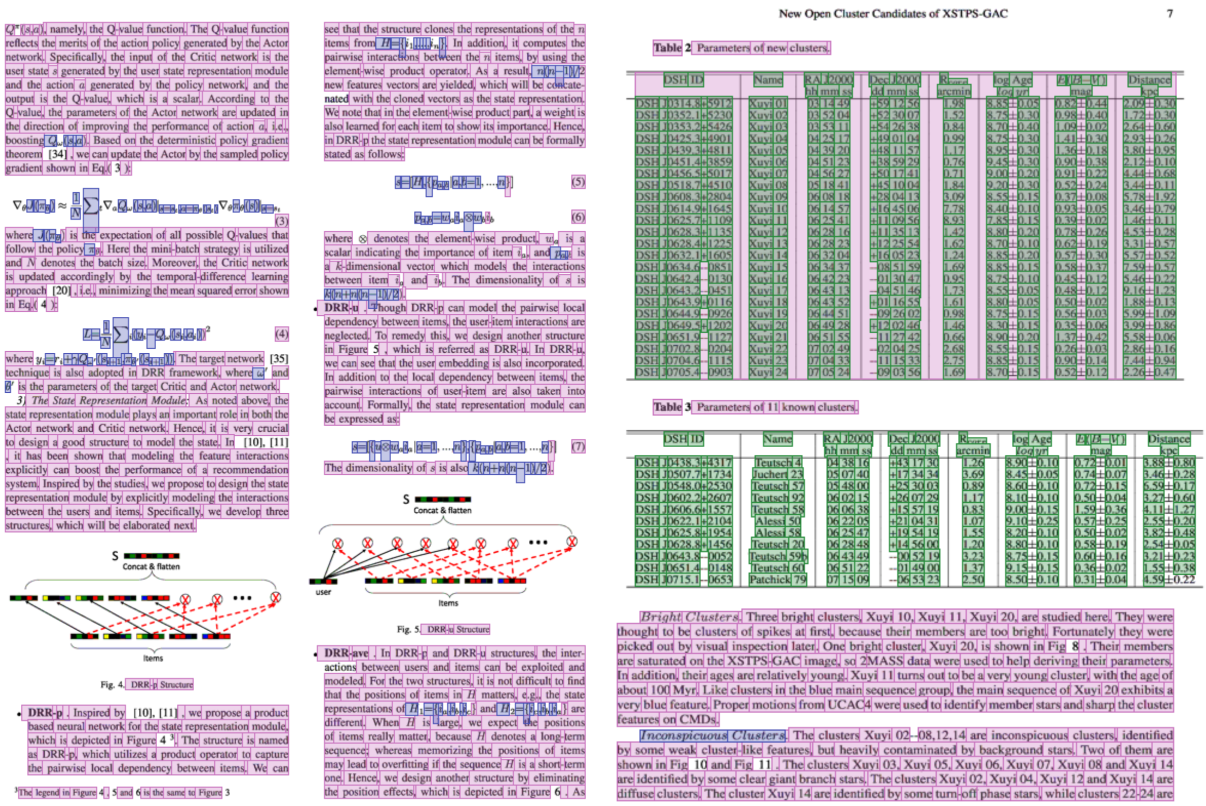


Figure A1: Dataset example. Bounding boxes of texts are highlighted in pink, mathematical expressions in blue, and tables in green.

B Output Examples

In Figure B1, we compared the output of LOCR and that of Nougat in Markdown format, together with the original PDF pages. Compared with Nougat, LOCR successfully handled the repetition problem. The corresponding part in PDF is highlighted in blue.

As a more clear illustration, Figure B2 shows the output of LOCR recompiled into PDF format.

Figure B3 shows the visualization of bounding boxes predicted by position detection head. LOCR predicts bounding boxes with high accuracy not only for plain texts, but also for figure captions, mathematical symbols and tables.

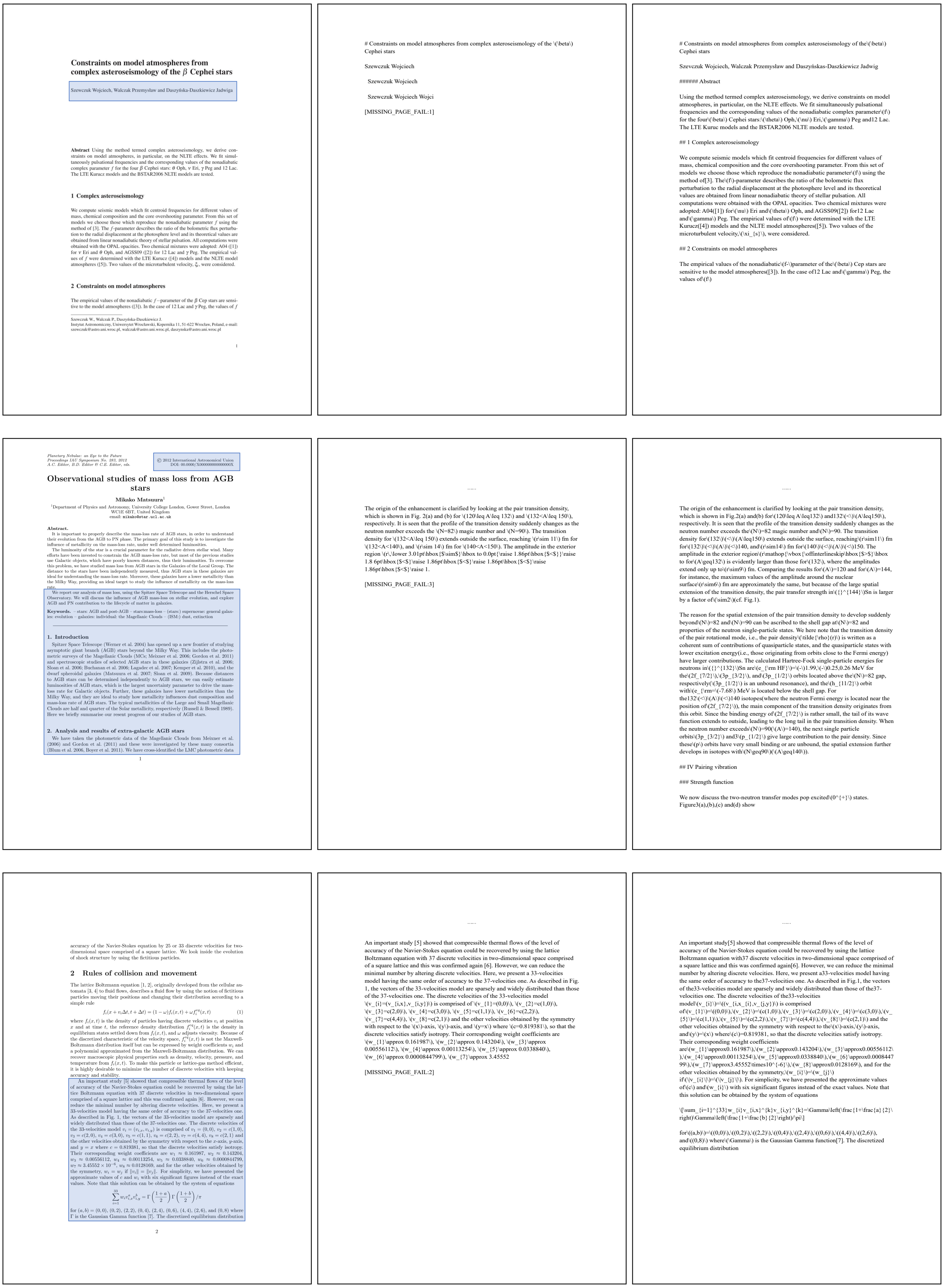


Figure B1: Examples of pages that Nougat failed to convert but LOCR succeeded. Left: Original PDF pages, with failed parts highlighted in blue. Medium: Markdown output by Nougat. Right: Markdown output by LOCR.

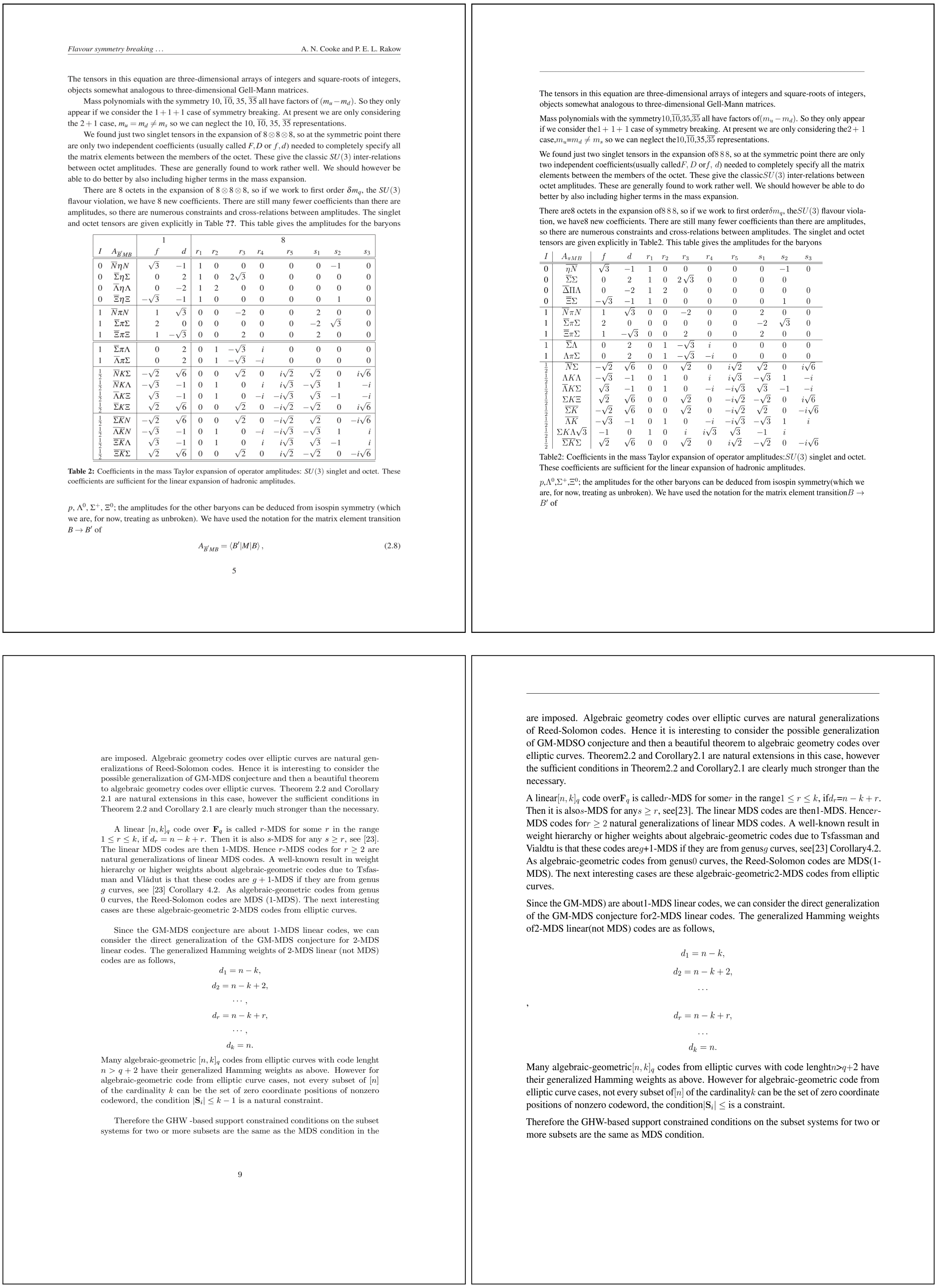


Figure B2: Examples of our model output. Left: Origin image of document page with tables and equations. Right: Model output converted to Markdown and rendered back into a PDF.

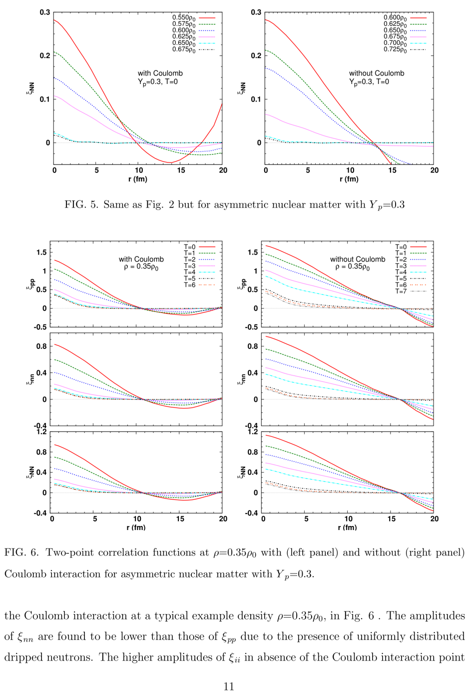


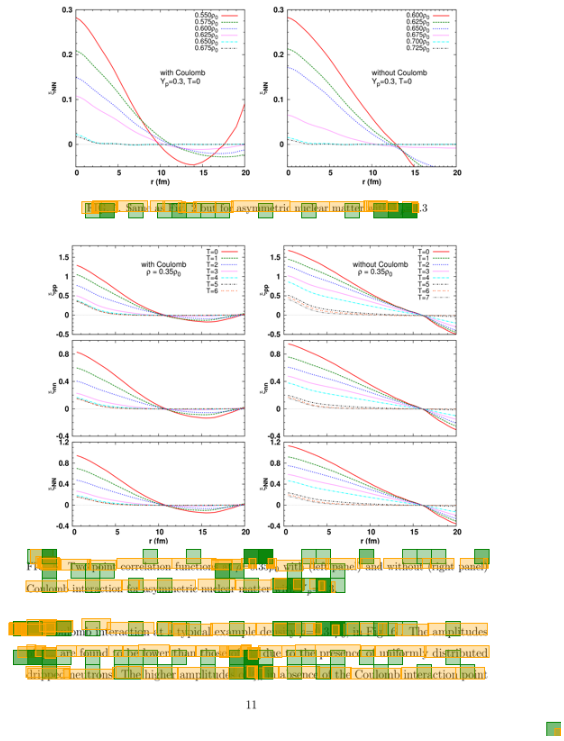
FIG. 5. Same as Fig. 2 but for asymmetric nuclear matter with Y_p=0.3

FIG. 6. Two-point correlation functions at ρ=0.35ρ_0 with (left panel) and without (right panel) Coulomb interaction for asymmetric nuclear matter with Y_p=0.3.

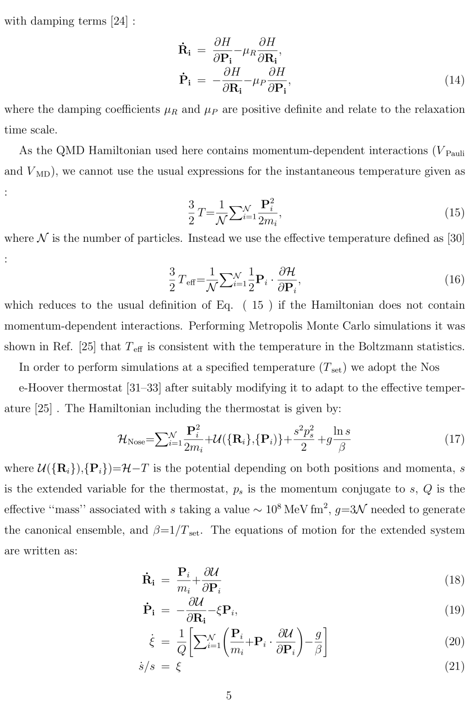
the Coulomb interaction at a typical example density ρ=0.35ρ_0, in Fig. 6 . The amplitudes of ξ_{nn} are found to be lower than those of ξ_{pp} due to the presence of uniformly distributed dripped neutrons. The higher amplitudes of ξ_{ii} in absence of the Coulomb interaction point

11

(a) Origin page with figures



(b) Result



with damping terms [24] :

$$\begin{aligned}\dot{\mathbf{R}}_\mathbf{i} &= \frac{\partial H}{\partial \mathbf{P_i}} - \mu_R \frac{\partial H}{\partial \mathbf{R_i}},\\ \dot{\mathbf{P}}_\mathbf{i} &= -\frac{\partial H}{\partial \mathbf{R_i}} - \mu_P \frac{\partial H}{\partial \mathbf{P_i}},\end{aligned} \tag{14}$$

where the damping coefficients μ_R and μ_P are positive definite and relate to the relaxation time scale.

As the QMD Hamiltonian used here contains momentum-dependent interactions (V_{Pauli} and V_{MD}), we cannot use the usual expressions for the instantaneous temperature given as :

$$\frac{3}{2}T = \frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}} \frac{\mathbf{P}_i^2}{2m_i}, \tag{15}$$

where $\mathcal{N}$ is the number of particles. Instead we use the effective temperature defined as [30] :

$$\frac{3}{2}T_{\text{eff}} = \frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}} \frac{1}{2}\mathbf{P}_i \cdot \frac{\partial \mathcal{H}}{\partial \mathbf{P}_i}, \tag{16}$$

which reduces to the usual definition of Eq. (15) if the Hamiltonian does not contain momentum-dependent interactions. Performing Metropolis Monte Carlo simulations it was shown in Ref. [25] that T_{eff} is consistent with the temperature in the Boltzmann statistics.

In order to perform simulations at a specified temperature (T_{set}) we adopt the Nos

e-Hoover thermostat [31–33] after suitably modifying it to adapt to the effective temperature [25] . The Hamiltonian including the thermostat is given by:

$$\mathcal{H}_{\text{Nose}} = \sum_{i=1}^{\mathcal{N}} \frac{\mathbf{P}_i^2}{2m_i} + \mathcal{U}(\{\mathbf{R}_i\},\{\mathbf{P}_i)\} + \frac{s^2 p_s^2}{2} + g\frac{\ln s}{\beta} \tag{17}$$

where $\mathcal{U}(\{\mathbf{R}_i\}),\{\mathbf{P}_i\}) = \mathcal{H} - T$ is the potential depending on both positions and momenta, s is the extended variable for the thermostat, p_s is the momentum conjugate to s, Q is the effective "mass" associated with s taking a value $\sim 10^8$ MeV fm^2, g=3$\mathcal{N}$ needed to generate the canonical ensemble, and β=1/T_{set}. The equations of motion for the extended system are written as:

$$\dot{\mathbf{R}}_\mathbf{i} = \frac{\mathbf{P}_i}{m_i} + \frac{\partial \mathcal{U}}{\partial \mathbf{P}_i} \tag{18}$$

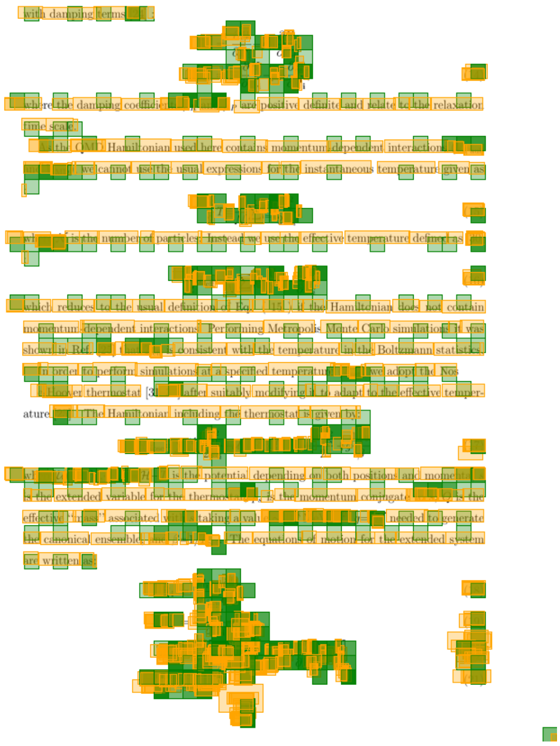
$$\dot{\mathbf{P}}_\mathbf{i} = -\frac{\partial \mathcal{U}}{\partial \mathbf{R_i}} - \xi \mathbf{P}_i, \tag{19}$$

$$\dot{\xi} = \frac{1}{Q}\left[\sum_{i=1}^{\mathcal{N}}\left(\frac{\mathbf{P}_i}{m_i} + \mathbf{P}_i \cdot \frac{\partial \mathcal{U}}{\partial \mathbf{P}_i}\right) - \frac{g}{\beta}\right] \tag{20}$$

$$\dot{s}/s = \xi \tag{21}$$

5

(c) Origin page with mathematical formulas



(d) Result

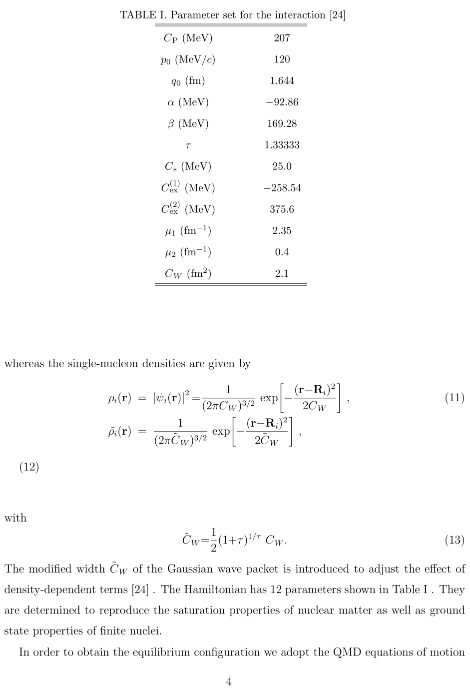


TABLE I. Parameter set for the interaction [24]

C_P (MeV)	207
p_0 (MeV/c)	120
q_0 (fm)	1.644
α (MeV)	−92.86
β (MeV)	169.28
τ	1.33333
C_s (MeV)	25.0
$C_{ex}^{(1)}$ (MeV)	−258.54
$C_{ex}^{(2)}$ (MeV)	375.6
μ_1 (fm^{-1})	2.35
μ_2 (fm^{-1})	0.4
C_W (fm^2)	2.1

whereas the single-nucleon densities are given by

$$\rho_i(\mathbf{r}) = |\psi_i(\mathbf{r})|^2 = \frac{1}{(2\pi C_W)^{3/2}} \exp\left[-\frac{(\mathbf{r}-\mathbf{R}_i)^2}{2C_W}\right], \quad (11)$$

$$\tilde{\rho}_i(\mathbf{r}) = \frac{1}{(2\pi \tilde{C}_W)^{3/2}} \exp\left[-\frac{(\mathbf{r}-\mathbf{R}_i)^2}{2\tilde{C}_W}\right],$$

(12)

with

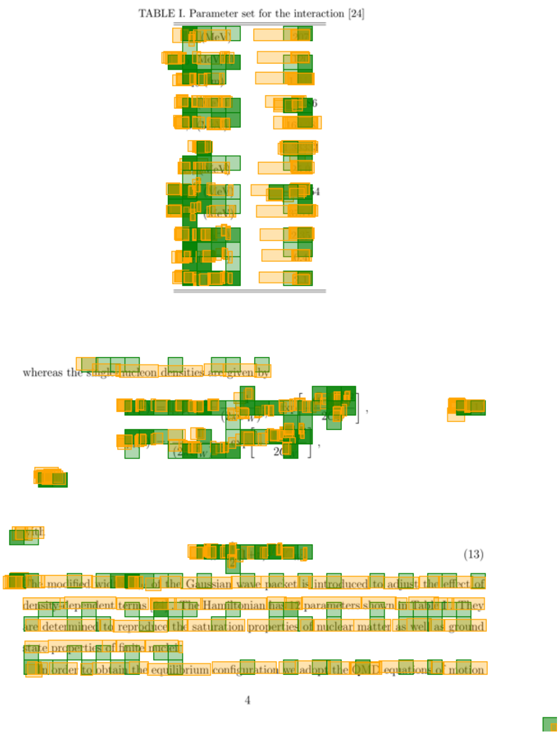
$$\tilde{C}_W = \frac{1}{2}(1+\tau)^{1/\tau}\, C_W. \quad (13)$$

The modified width $\tilde{C}_W$ of the Gaussian wave packet is introduced to adjust the effect of density-dependent terms [24] . The Hamiltonian has 12 parameters shown in Table I . They are determined to reproduce the saturation properties of nuclear matter as well as ground state properties of finite nuclei.

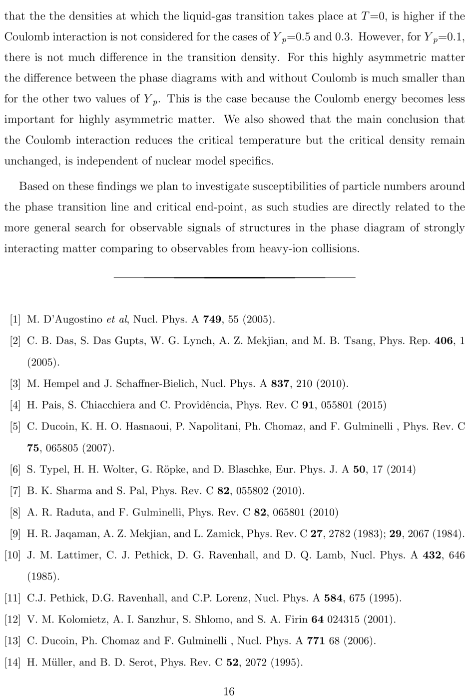
In order to obtain the equilibrium configuration we adopt the QMD equations of motion

4

(e) Origin page with tables



(f) Result



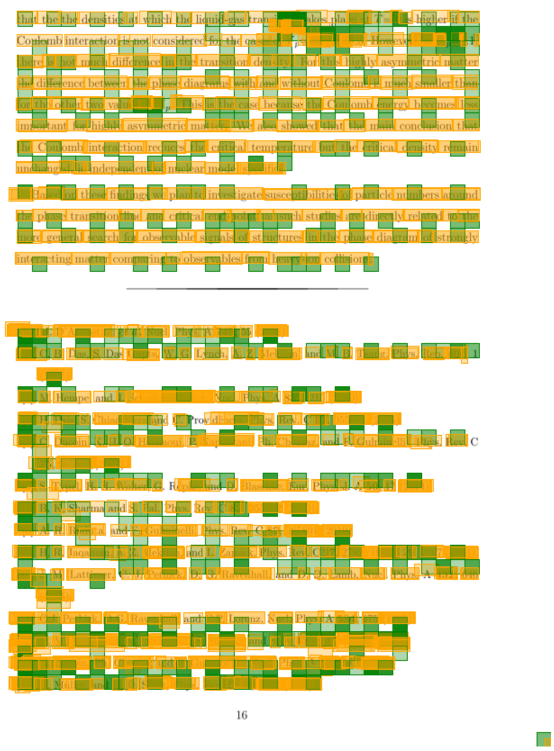
that the the densities at which the liquid-gas transition takes place at T=0, is higher if the Coulomb interaction is not considered for the cases of Y_p=0.5 and 0.3. However, for Y_p=0.1, there is not much difference in the transition density. For this highly asymmetric matter the difference between the phase diagrams with and without Coulomb is much smaller than for the other two values of Y_p. This is the case because the Coulomb energy becomes less important for highly asymmetric matter. We also showed that the main conclusion that the Coulomb interaction reduces the critical temperature but the critical density remain unchanged, is independent of nuclear model specifics.

Based on these findings we plan to investigate susceptibilities of particle numbers around the phase transition line and critical end-point, as such studies are directly related to the more general search for observable signals of structures in the phase diagram of strongly interacting matter comparing to observables from heavy-ion collisions.

[1] M. D'Augostino *et al*, Nucl. Phys. A **749**, 55 (2005).
[2] C. B. Das, S. Das Gupts, W. G. Lynch, A. Z. Mekjian, and M. B. Tsang, Phys. Rep. **406**, 1 (2005).
[3] M. Hempel and J. Schaffner-Bielich, Nucl. Phys. A **837**, 210 (2010).
[4] H. Pais, S. Chiacchiera and C. Providência, Phys. Rev. C **91**, 055801 (2015)
[5] C. Ducoin, K. H. O. Hasnaoui, P. Napolitani, Ph. Chomaz, and F. Gulminelli , Phys. Rev. C **75**, 065805 (2007).
[6] S. Typel, H. H. Wolter, G. Röpke, and D. Blaschke, Eur. Phys. J. A **50**, 17 (2014)
[7] B. K. Sharma and S. Pal, Phys. Rev. C **82**, 055802 (2010).
[8] A. R. Raduta, and F. Gulminelli, Phys. Rev. C **82**, 065801 (2010)
[9] H. R. Jaqaman, A. Z. Mekjian, and L. Zamick, Phys. Rev. C **27**, 2782 (1983); **29**, 2067 (1984).
[10] J. M. Lattimer, C. J. Pethick, D. G. Ravenhall, and D. Q. Lamb, Nucl. Phys. A **432**, 646 (1985).
[11] C.J. Pethick, D.G. Ravenhall, and C.P. Lorenz, Nucl. Phys. A **584**, 675 (1995).
[12] V. M. Kolomietz, A. I. Sanzhur, S. Shlomo, and S. A. Firin **64** 024315 (2001).
[13] C. Ducoin, Ph. Chomaz and F. Gulminelli , Nucl. Phys. A **771** 68 (2006).
[14] H. Müller, and B. D. Serot, Phys. Rev. C **52**, 2072 (1995).

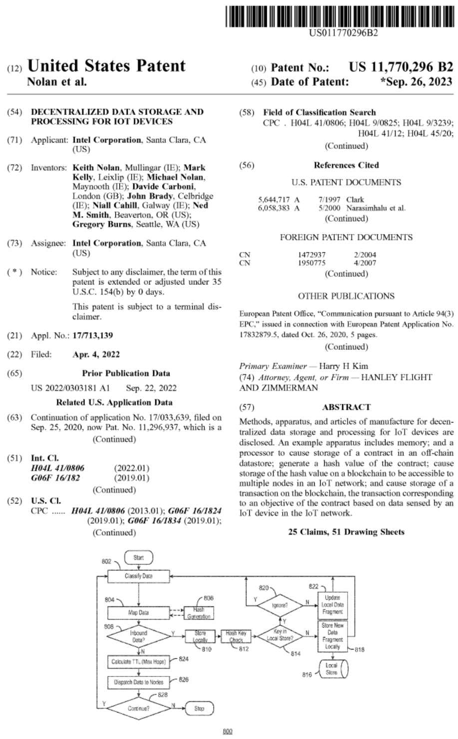
16

(g) Origin page with references

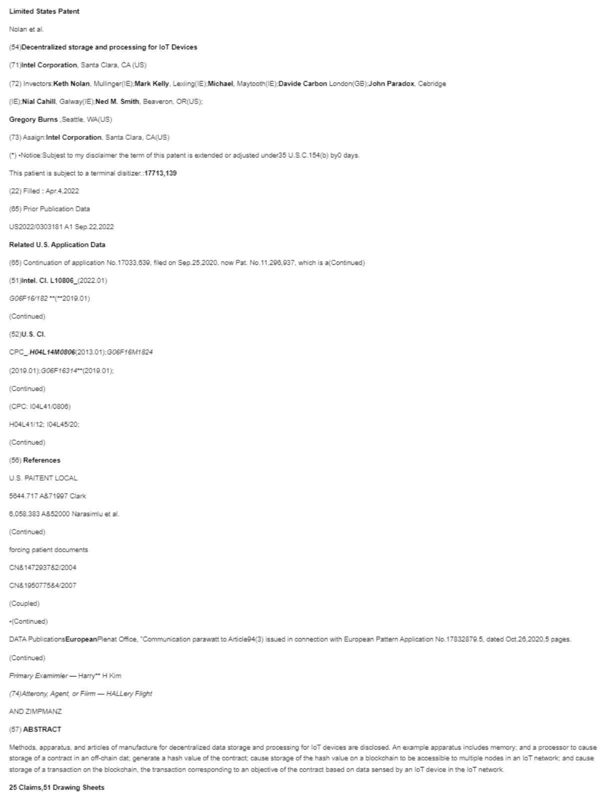


(h) Result

Figure B3: Example of position prediction. Green box: Rough result of grid classification. Yellow: Final result of box regression.



(i) Origin patent page

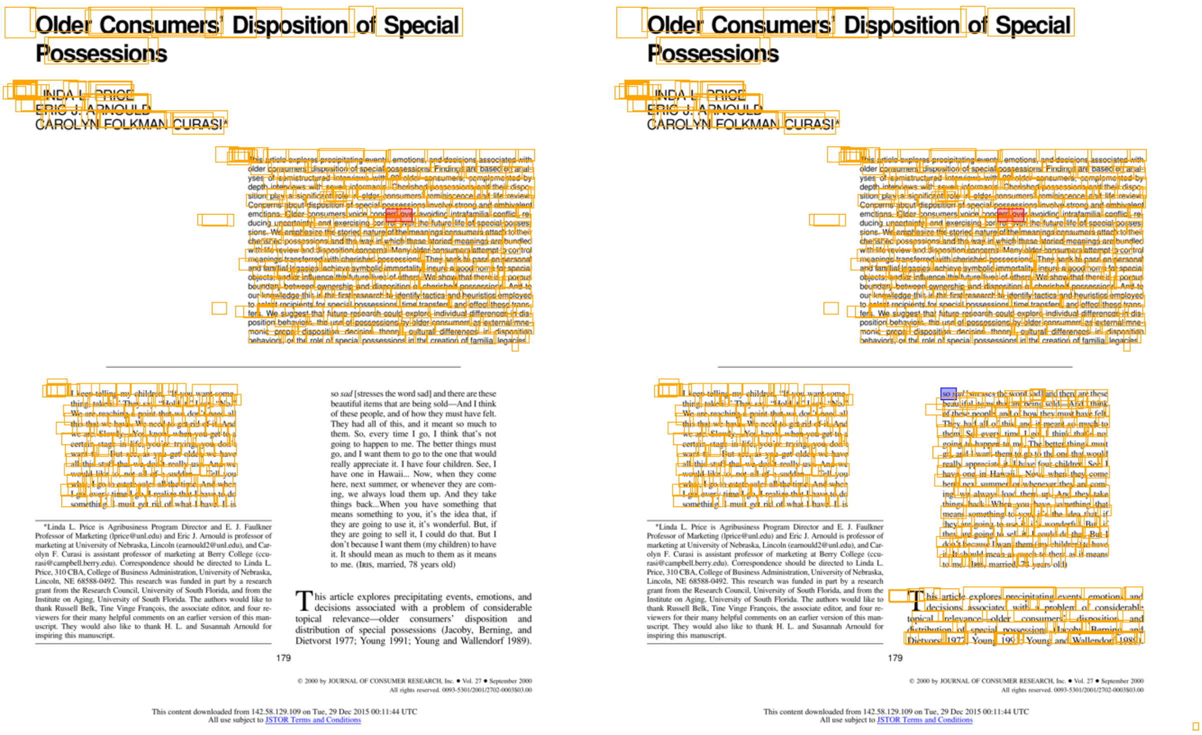


(j) Result

Figure B4: Example of our model output on patent documents. LOCR is able to parse a broader range of layouts and document domains beyond academic papers, indicating the flexibility of location-based OCR method. Besides, with the interactive mode and the model automatically predicting positions, minimal human intervention is required to acquire additional out-of-domain data, particularly the positional bounding box labels. This paves the way for broader applications of location-based OCR method.

C Interactive Mode

Figure C1 shows the interactive process with human intervention. The orange bounding boxes denote the areas that have been scanned by the model. The model predicted a low confidence score when it decoded to the position shown in 1(a), with the incorrectly predicted position highlighted in red. In 1(c), human gave a box prompt highlighted in blue and the model output the subsequent contents smoothly and correctly.



(k) A case model predicting wrong position (l) Result

Figure C1: Visualization of interaction on out-of-domain documents. Red box: Wrong position. Blue box: Human prompt input.

D Statistics of Test Documents

As a complementary illustration for Table 2, we show the histograms of the number of pages per document in Figure D1. Consistent with the conclusion in Table 2, when counting in document number, domains with more pages per document, such as marketing, have a higher generation failure rate.

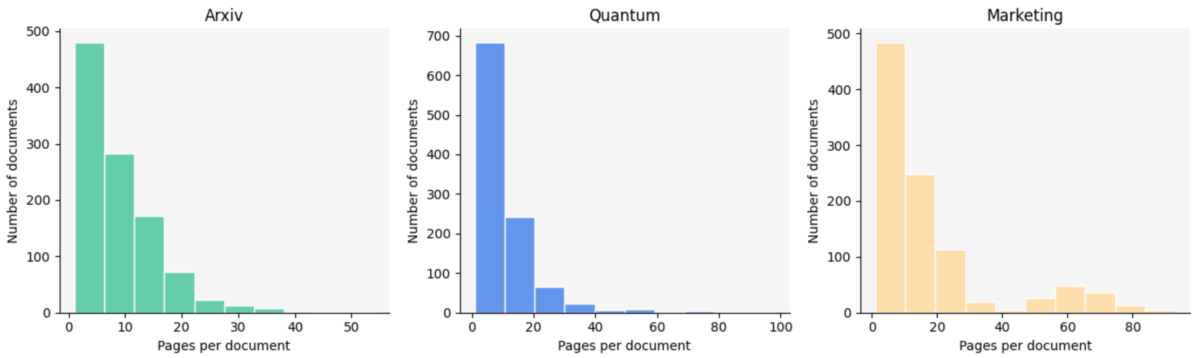
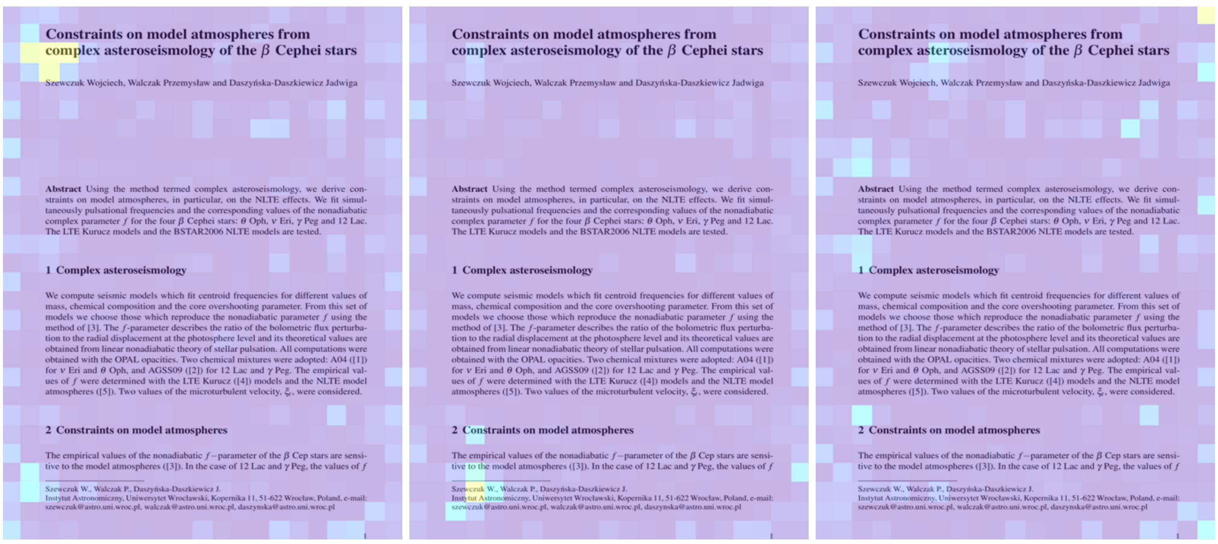


Figure D1: Histograms of the number of pages per document in each repetition test set.

E A case when Nougat gets trapped into repetition

Figure E1 shows a case when nougat got trapped into repetition. After decoding the name of the first author, Nougat tried to find the correlation between the footnote and the authors but failed. The heatmap of cross-attenions ended with cycling through the three subfigures and the output ended with repeating the name "Szewczuk Wojciech Szewczuk Wojciech Szewczuk Wojciech Wojci". The original PDF page, the output of Nougat and that of LOCR is shown in Figure B1.



(a) Correct attentions for the authors. (b) Correct attentions for the footnote (c) Incorrect attentions when repetition.

Figure E1: The heatmap of cross-attention of Nougat, in which yellow denotes larger attention scores and purple denotes smaller scores. Left: Cross-attention scores when Nougat decoded to the name of the first author. Medium: Cross-attention scores when Nougat tried to decode the footnote. Right: Cross-attention scores when Nougat began repetition and failed to find the correct position.