

## СОДЕРЖАНИЕ ПРЕЗЕНТАЦИИ

- ЦЕЛЬ И АКТУАЛЬНОСТЬ MINERU
- СУЩЕСТВУЮЩИЕ ПОДХОДЫ К ИЗВЛЕЧЕНИЮ ИЗ PDF-ДОКУМЕНТОВ
- АРХИТЕКТУРА MINERU: ЭТАПЫ ПРЕДОБРАБОТКИ, ПАРСИНГА, ПОСТОБРАБОТКИ И КОНВЕРТАЦИИ
- ПРАКТИЧЕСКАЯ ЧАСТЬ: ПРИМЕРЫ ТЕСТОВЫХ ДОКУМЕНТОВ И РЕЗУЛЬТАТЫ

# MINERU

ОТКРЫТОЕ РЕШЕНИЕ ДЛЯ  
ТОЧНОГО ИЗВЛЕЧЕНИЯ  
СОДЕРЖИМОГО ДОКУМЕНТОВ



## MINERU

---

– инструмент для высокоточного извлечения содержимого из PDF-документов

## АКТУАЛЬНОСТЬ

---

- Необходимость извлечения данных усилилась ростом LLM – веб-данных уже мало, документы содержат ценные знания

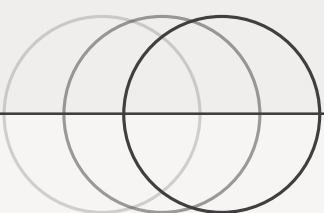
## ЦЕЛЬ

---

– Обилие форматов и элементов (текст, формулы, таблицы, изображения) делает задачу сложной

# Существующие подходы

К ИЗВЛЕЧЕНИЮ ИЗ PDF-ДОКУМЕНТОВ



## OCR-МЕТОДЫ

применяют распознавание текста по изображению. Хорошо работают для простых текстовых страниц, но дают ошибки при наличии таблиц, формул, рисунков

## БИБЛИОТЕЧНЫЙ ПАРСИНГ

используют встроенные возможности для извлечения текста без OCR. Быстрее и точнее для обычных текстовых PDF, но не справляются с графикой, формулами, сложной вёрсткой

## МУЛЬТИ-МОДУЛЬНЫЕ

сначала детектируют лэйаут, затем в каждом блоке применяют соответствующие модели. В теории этот подход точен, но существующие решения чаще заточены на научные статьи

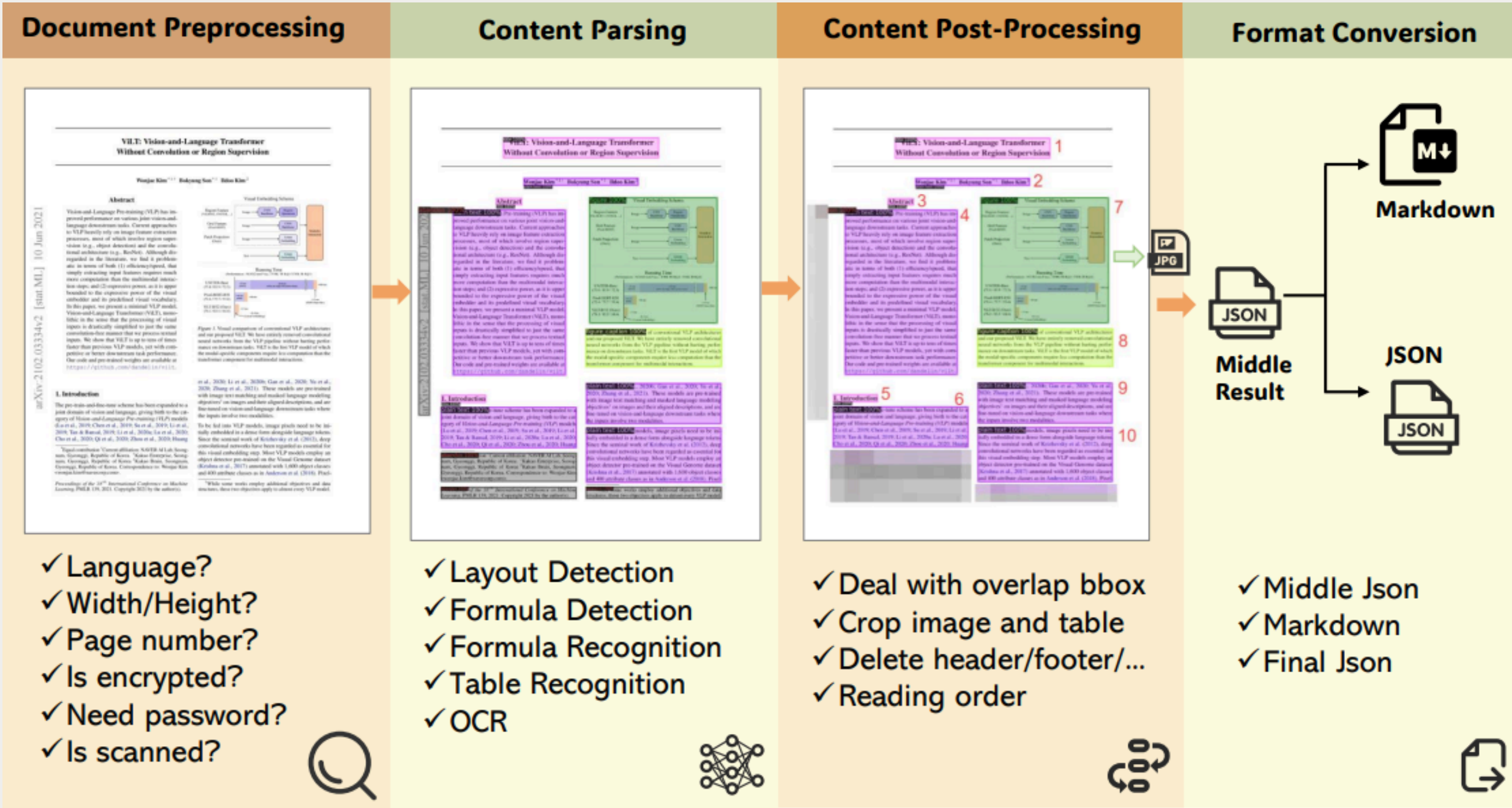
## MULTIMODAL LLM

целиком обрабатывают изображение страницы с помощью мультимодальных нейросетей. Имеют проблемы с разнообразием данных и высокой вычислительной сложностью

ПРЕДОБРАБОТКА  
ПАРСИНГ  
ПОСТОБРАБОТКА  
КОНВЕРТАЦИЯ



# Архитектура



# Предобработка и парсинг



На первом этапе MinerU:

- проверяет файл (PDF ли это, не зашифрован ли, требуется пароль)
- определяет количество страниц
- язык
- обнаруживает, является ли документ отсканированным или «цифровым»

**Это важно:** отсканированные PDF нужно обрабатывать через OCR, а текстовые – можно прочитать напрямую

MinerU также вычисляет размеры страниц и фиксирует метаданные

- Сначала выполняется layout-анализ
- Параллельно обнаруживаются формулы
- Затем для разных областей применяются специализированные распознаватели:
  - OCR (PaddleOCR) для текстовых блоков
  - UniMERNet для распознавания формул
  - Табличные модели (TableMaster и StructEqTable) для извлечения данных из таблиц

# Постобработка и конвертация



После получения результатов моделей MinerU «чистит» и упорядочивает содержимое:

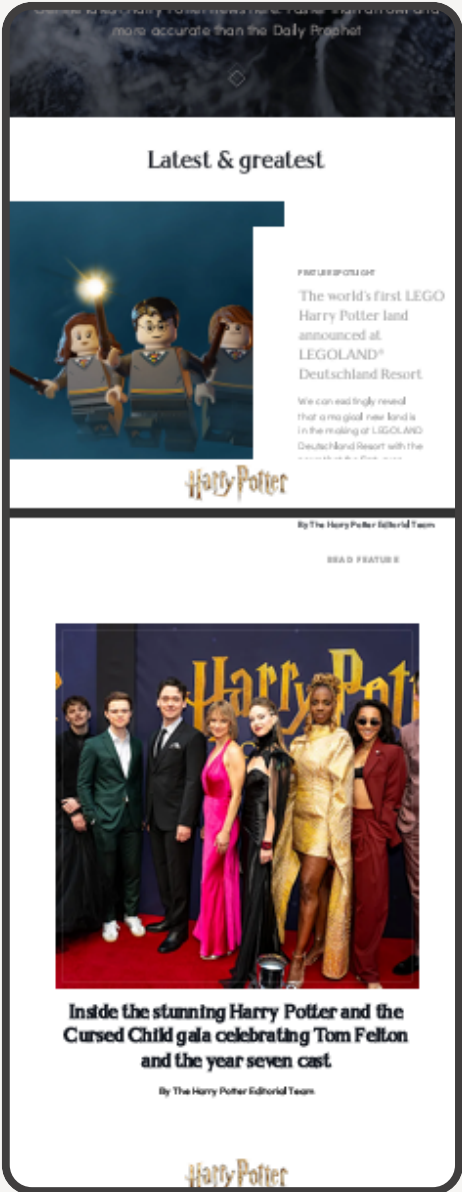
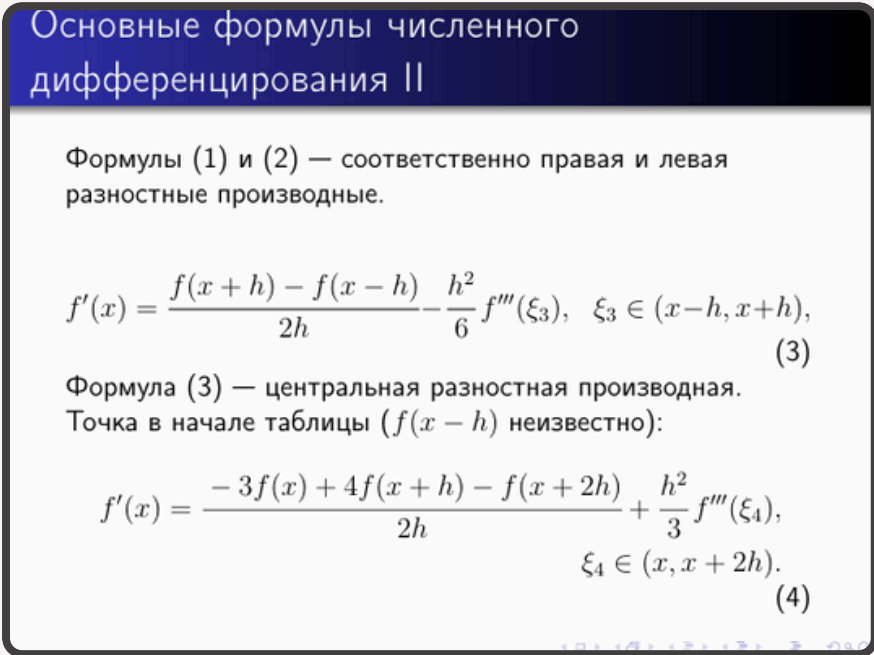
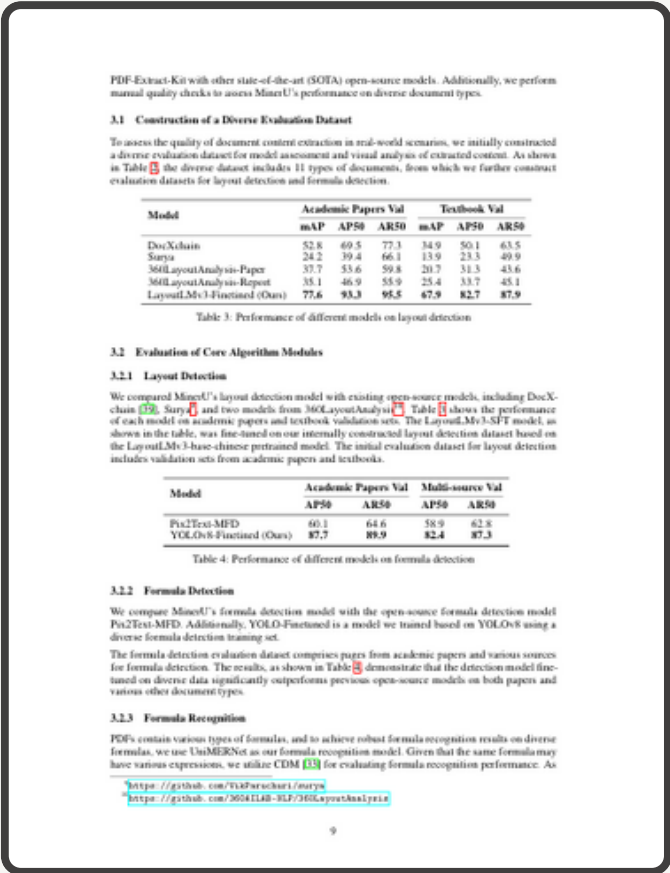
- Устраняются пересечения блоков
- Удаляются дубли
- Страницы разбиваются на группы по читаемым колонкам
- Фильтруются ненужные области для повышения читабельности
- Тексты абзацев при необходимости склеиваются с помощью правил

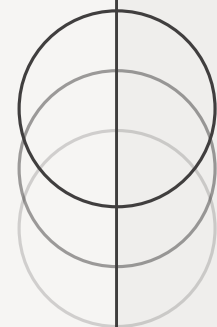
Окончательный результат представляется в формате JSON/Markdown

MinerU сохраняет

- Промежуточную структуру JSON: упорядоченный массив «para\_blocks», где каждый элемент – сегмент контента.
- Markdown или финальный JSON (с учетом вырезанных изображений/таблиц и пр.).

# Обзор тестовых данных





ПРАКТИЧЕСКАЯ ЧАСТЬ

# Замеченные плюсы модели

Exclusive artwork reveal and Q&A<sup>11</sup>  
with Harry Potter and the Goblet of  
Fire Interactive Edition illustrator Karl

James Mountford<sup>12</sup>

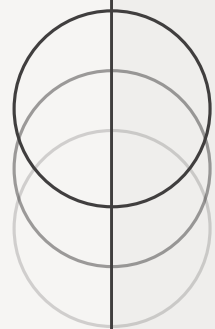
By The Harry Potter Editorial Team<sup>13</sup>



By The Harry Potter Editorial Team<sup>1</sup>

Take a trip around the wizarding<sup>2</sup>  
world with Harry Potter: Hogwarts  
Mystery

- Пропуск не важных - повторяющихся элементов на веб-странице



# Замеченные плюсы модели

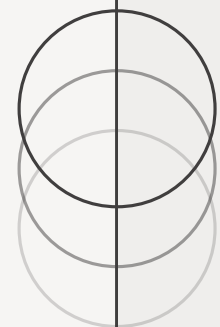
Таблица 1<sup>1</sup><sup>2</sup>

$x_k$	$y_k$	$f'_T$	$\tilde{f}'$ $O(h^2)$	погр. $O(h^2)$	$\tilde{f}'$ $O(h^4)$	погр. $O(h^4)$	$f''_T$	$\tilde{f}''$ $O(h^2)$	погр. $O(h^2)$
$x_0$	$y_0$		(4)		(7)			(12)	
$x_1$	$y_1$		(3)		(8)			(6)	
$x_2$	$y_2$		(3)		(9)			(6)	
...	...		...		...			...	
$x_{m-2}$	$y_{m-2}$		(3)		(9)			(6)	
$x_{m-1}$	$y_{m-1}$		(3)		(10)			(6)	
$x_m$	$y_m$		(5)		(11)			(13)	

Этап анализа результатов.<sup>3</sup>

Предложение выбрать функцию, создать новую таблицу и <sup>4</sup> провести новые расчеты или выйти из программы.

- Пропуск не важных - повторяющихся элементов на веб-странице
- Распознавание таблиц



# Замеченные плюсы модели

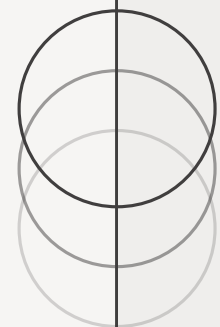


И доктор пришил ему ножки,<sup>2</sup>  
И зайчика прыгает снова.

А с ним и зайчиха-мать<sup>3</sup>  
Тоже пошла танцевать.  
И смеется она и кричит:  
„Ну, спасибо тебе, Айболит!“



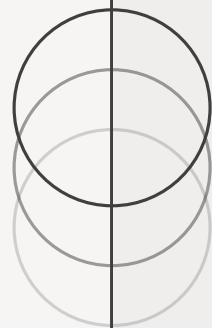
- Пропуск не важных - повторяющихся элементов на веб-странице
- Распознавание таблиц
- Корректное выделение текста от иллюстраций даже на сканированных документах



# Замеченные плюсы модели



- Пропуск не важных - повторяющихся элементов на веб-странице
- Распознавание таблиц
- Корректное выделение текста от иллюстраций даже на сканированных документах
- Верный алгоритм последовательного чтения даже на сложных шаблонах



# Замеченные плюсы модели

PDF-Extract-Kit with other state-of-the-art (SOTA) open-source models. Additionally, we perform manual quality checks to assess MinerU's performance on diverse document types.

3.1 Construction of a Diverse Evaluation Dataset

To assess the quality of document content extraction in real-world scenarios, we initially constructed a diverse evaluation dataset for model assessment and visual analysis of extracted content. As shown in Table 3, the diverse dataset includes 11 types of documents, from which we further construct evaluation datasets for layout detection and formula detection.

Model	Academic Papers Val			Textbook Val		
	mAP	AP50	AR50	mAP	AP50	AR50
DocXchain	52.8	69.5	77.3	34.9	50.1	63.5
Surya	24.2	39.4	66.1	13.9	23.3	49.9
360LayoutAnalysis-Paper	37.7	53.6	59.8	20.7	31.3	43.6
360LayoutAnalysis-Report	35.1	46.9	55.9	25.4	33.7	45.1
LayoutLMv3-Finetuned (Ours)	77.6	93.3	95.5	67.9	82.7	87.9

Table 3: Performance of different models on layout detection

3.2 Evaluation of Core Algorithm Modules

3.2.1 Layout Detection

We compared MinerU's layout detection model with existing open-source models, including DocXchain, Surya, and two models from 360LayoutAnalysis. Table 3 shows the performance of each model on academic papers and textbook validation sets. The LayoutLMv3-SFT model, as shown in the table, was fine-tuned on our internally constructed layout detection dataset based on the LayoutLMv3-base-chinese pretrained model. The initial evaluation dataset for layout detection includes validation sets from academic papers and textbooks.

Model	Academic Papers Val		Multi-source Val	
	AP50	AR50	AP50	AR50
Pix2Text-MFD	60.1	64.6	58.9	62.8
YOLOv8-Finetuned (Ours)	87.7	89.9	82.4	87.3

Table 4: Performance of different models on formula detection

3.2.2 Formula Detection

We compare MinerU's formula detection model with the open-source formula detection model Pix2Text-MFD. Additionally, YOLO-Finetuned is a model we trained based on YOLOv8 using a diverse formula detection training set.

The formula detection evaluation dataset comprises pages from academic papers and various sources for formula detection. The results, as shown in Table 4, demonstrate that the detection model fine-tuned on diverse data significantly outperforms previous open-source models on both papers and various other document types.

3.2.3 Formula Recognition

PDFs contain various types of formulas, and to achieve robust formula recognition results on diverse formulas, we use UniMERNet as our formula recognition model. Given that the same formula may have various expressions, we utilize CDM for evaluating formula recognition performance. As

<https://github.com/VikParuchuri/surya>

<https://github.com/360AILAB-NLP/360LayoutAnalysis>

- Пропуск не важных - повторяющихся элементов на веб-странице
- Распознавание таблиц
- Кореектное выделение текста от иллюстраций даже на сканированных документах
- Верный алгоритм последовательного чтения даже на сложных шаблонах
- Работа со сложной структурой данных, соблюдение иерархии и верного расположения вложений



# Ошибки модели

СЛОЖНОСТИ В РАБОТЕ  
С ДЛИННЫМИ И  
ГРОМОЗКИМИ  
ФОРМУЛАМИ

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{6} f'''(\xi_3), \quad \xi_3 \in (x-h, x+h), \quad (3)$$

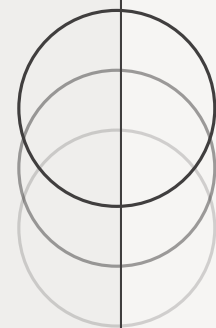
Формула (3) — центральная разностная производная.  
Точка в начале таблицы ( $f(x-h)$  неизвестно):

$$f'(x) = \frac{-3f(x) + 4f(x+h) - f(x+2h)}{2h} + \frac{h^2}{3} f'''(\xi_4), \quad \xi_4 \in (x, x+2h). \quad (4)$$

При использовании формулы (4) в результате о допустимых в каждом значении функции и не превосходящих по модулю  $\varepsilon$ , оценка для суммар погрешности будет выглядеть следующим образом

$$|R_\varepsilon(x, h, f)| \leq \frac{8\varepsilon}{2h} + \frac{h^2}{3} M_3, \quad M_3 = \max |f'''(\xi)|, \quad \xi \in [x, x+2h]$$

Оптимальный шаг, т. е. такой, при котором обеспечена минимальная суммарная погрешность, находится, как решение задачи на экстремум.



# Ошибки модели

Нр р АТ! ОН . М МК, 1 В, WMeNBenua!

Всех , нТ А!

# 2

а а: ,0, Ме уКуС оса!"

А 6, К В!"

W Закраа:,Аñ-а!

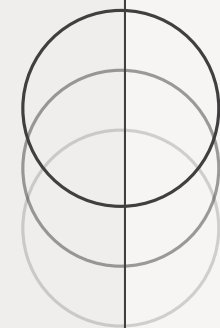
!

!

О е, W еу перерао НК, 6 , Ма 3 .

СЛОЖНОСТИ В РАБОТЕ  
С ДЛИННЫМИ И  
ГРОМОЗКИМИ  
ФОРМУЛАМИ

НЕКОРРЕКТНОЕ  
РАСПОЗНОВАНИЕ  
ТЕКСТА СО СКАНА ДЛЯ  
РУССКОГО ЯЗЫКА



# Ошибки модели

СЛОЖНОСТИ В РАБОТЕ  
С ДЛИННЫМИ И  
ГРОМОЗКИМИ  
ФОРМУЛАМИ

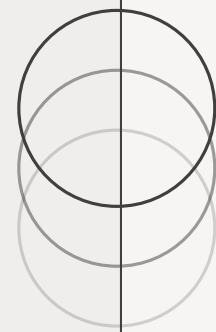
НЕКОРРЕКТНОЕ  
РАСПОЗНОВАНИЕ  
ТЕКСТА СО СКАНА ДЛЯ  
РУССКОГО ЯЗЫКА

ПРОПУСК ВАЖНЫХ  
ЭЛЕМЕНТОВ

“  
Будущее принадлежит  
тем, кто верит в красоту  
своих мечтаний.

— Элеонора Рузвельт

— Элеонора Рузвельт



# Ошибки модели

FEATURE SPOTLIGHT 5

The world's first LEGO<sup>®</sup>  
Harry Potter land  
announced at  
LEGOLAND<sup>®</sup>  
Deutschland Resort

We can excitingly reveal  
that a magical new land is  
in the making at LEGOLAND  
Deutschland Resort with the

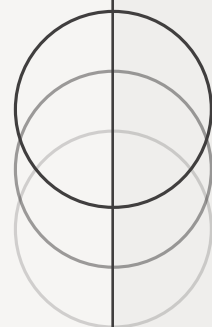
fter

СЛОЖНОСТИ В РАБОТЕ  
С ДЛИННЫМИ И  
ГРОМОЗКИМИ  
ФОРМУЛАМИ

НЕКОРРЕКТНОЕ  
РАСПОЗНОВАНИЕ  
ТЕКСТА СО СКАНА ДЛЯ  
РУССКОГО ЯЗЫКА

ПРОПУСК ВАЖНЫХ  
ЭЛЕМЕНТОВ

ПУТАНИЦА В ПОРЯДКЕ  
БЛОКОВ ЭЛЕМЕНТОВ



## СОДЕРЖАНИЕ ПРЕЗЕНТАЦИИ

- ЦЕЛЬ И АКТУАЛЬНОСТЬ MINERU
- СУЩЕСТВУЮЩИЕ ПОДХОДЫ К ИЗВЛЕЧЕНИЮ ИЗ PDF-ДОКУМЕНТОВ
- АРХИТЕКТУРА MINERU: ЭТАПЫ ПРЕДОБРАБОТКИ, ПАРСИНГА, ПОСТОБРАБОТКИ И КОНВЕРТАЦИИ
- ПРАКТИЧЕСКАЯ ЧАСТЬ: ПРИМЕРЫ ТЕСТОВЫХ ДОКУМЕНТОВ И РЕЗУЛЬТАТЫ

# MINERU

ОТКРЫТОЕ РЕШЕНИЕ ДЛЯ  
ТОЧНОГО ИЗВЛЕЧЕНИЯ  
СОДЕРЖИМОГО ДОКУМЕНТОВ