INTERNATIONAL CONFERENCE ON EDUCATIONAL TECHNOLOGY AND ARTIFICIAL INTELLIGENCE

# SmartDoc

*A Client-Side Paginated Web Platform for Word-to-Interactive-Textbook Conversion with In-DOM AI Tutoring and Asynchronous Reading Analytics*

**Nguyễn Đức Thuận**

Faculty of Information Technology · National Economics University (NEU), Hanoi, Vietnam

# What we'll cover

# A structural gap in the smart-classroom pipeline



W

## What instructors use

Microsoft Word — already familiar, already trusted.

</>

## What students need

Live, analytics-rich web documents.

## The missing bridge

- Static PDF/EPUB — no live DOM for AI or analytics.
- Server-side PDF — collapses under classroom-scale concurrency.
- Generic chatbots — retrieve indiscriminately, prone to hallucination.

# Where the literature stops — and SmartDoc starts



## 1 Paged media

Headless Chrome renders PDFs — no live DOM, heavy server CPU.

## 2 Context-aware AI assistants

RAG helps, but readers usually retrieve across the whole corpus.

## 3 Learning analytics

LMS clickstreams are coarse; reader-level behavior is rarely captured.

## 4 Smart-classroom infrastructure

A gap between research prototypes and deployable infrastructure.

# SmartDoc: a paginated yet live DOM as the unifying primitive



1

### Automated DOM-cleansing pipeline

Handles messy Word→ JSON structures with zero manual intervention.

2

### Client-side pagination

Fragmentation runs in the browser — the server just streams JSON.

3

### Page-bounded, self-hosted AI tutor

Retrieval scope bound to the open document; self-hosted LLM.

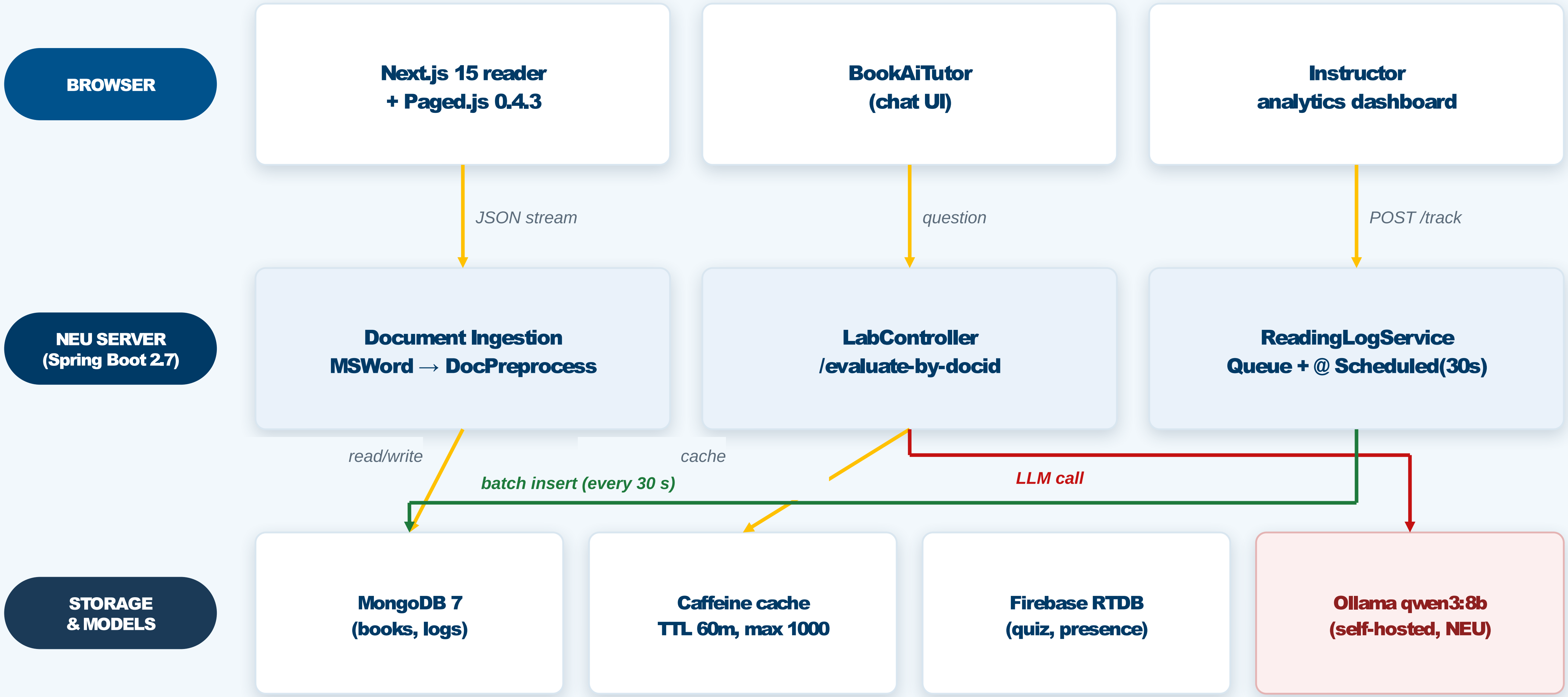
4

### Non-blocking event-logging

A queue drained every 30s captures six reading-event classes.

# Three server-side subsystems on a single Spring Boot backend





# Technology and version of the NEU deployment



*Every tier is swappable via environment variables.*

**FRONTEND**

- **Next.js** 15.1.9
- **React** 18.3.1
- **Paged.js** 0.4.3

**BACKEND**

- **Spring Boot** Java 11 · 2.7.12
- **Apache POI** 4.1.1
- **OkHttp** 4.12.0

**STORAGE**

- **MongoDB** v7 · driver 4.7.1

**CACHE**

- **Caffeine** TTL 60m · max 1,000

**LLM**

- **Ollama qwen3:8b** self-hosted · NEU

# Why a cleansing pipeline?



*Three pathologies dominate raw .docx output:*

**1**

### Fragmented text runs

Style changes fragment one paragraph into many runs.

**2**

### Vendor style names

Headings pass through as raw HEADING_1–5 style names.

**3**

### Office Math (OMML)

Math content must be transcoded to MathML.

# Five-pass DOM-cleansing pipeline



*Five idempotent passes turn raw .docx into paginate-ready JSON.*

1. **Style-to-semantic mapping**
   Word styles → semantic format (TITLE, HEADING).
2. **Text-run merging**
   Identical-style runs merged per paragraph.
3. **Math transcoding**
   OMML transcoded to MathML.
4. **Structural extraction**
   Paragraphs, headings, lists, tables extracted.
5. **Identifier injection**
   Stable IDs injected for AI + analytics.

**Input:** raw .docx **Output:** clean JSON + HTML, paginate-ready

# Client-side pagination as the scalability primitive



*Paged.js measures and fragments content entirely in the browser.*

**SERVER-SIDE PDF COMPILER**

**~1 GB**

per render request

✗ 100 concurrent students exhaust one server.

**SMARTDOC — CLIENT-SIDE PAGED.JS**

**~5.1 GB**

total, backend+frontend+Mongo

✓ Server just streams JSON — ample headroom remains.

# Page-bounded AI tutor, backed by a self-hosted LLM



**Student reads a page** → **documentId resolved** → **POST to /evaluate-by-docid** → **Retrieval scoped to that document** → **Ollama qwen3:8b (self-hosted, NEU)**

### Sovereign, not vendor-locked

No query leaves the university network.

### Layout-aware retrieval scope

Scope follows the open document.

### Quick-action templates

Generated from the table of contents.

# Non-blocking, six-channel reading analytics



**LEFT_CLICK**

*Spatial focus*

**TEXT_SELECTION**

*Copy intent*

**LONG_SCROLL**

*Skim vs. deep-read*

**CHANGE_CHAPTER**

*Study path*

**TOGGLE_AI_CHAT**

*Help-seeking*

**CLICK_TOC**

*Navigation strategy*

*@PreDestroy flushes residual logs — no data lost on restart.*

## NON-BLOCKING WRITE PATH

**Browser**
**POST /track**

↓

**LabController**
**HTTP 202 immediately**

↓

**ConcurrentLinkedQueue**
**(in RAM)**

↓

**@Scheduled(30s)**
**bulk insert**

↓

**MongoDB**
**reading_logs**

# Instructor analytics dashboard



*A first quantitative window into reading behavior.*

### 1 Per-student & per-chapter counts

Event volume per student, per chapter.

### 2 Most-visited chapters

Which sections draw the most repeat traffic.

### 3 Longest text selections

“What did students copy?” — concept signals.

### 4 Most-clicked TOC headings

Which entries students jump to directly.

# Deployed at FIT, NEU — real production load



**5**
course textbooks in production

**1,500**
Word pages ingested

**248,036**
words (91.3 MB total)

**4,946**
real reading events captured

**~5.1 GB**
combined RAM (backend+frontend+Mongo)

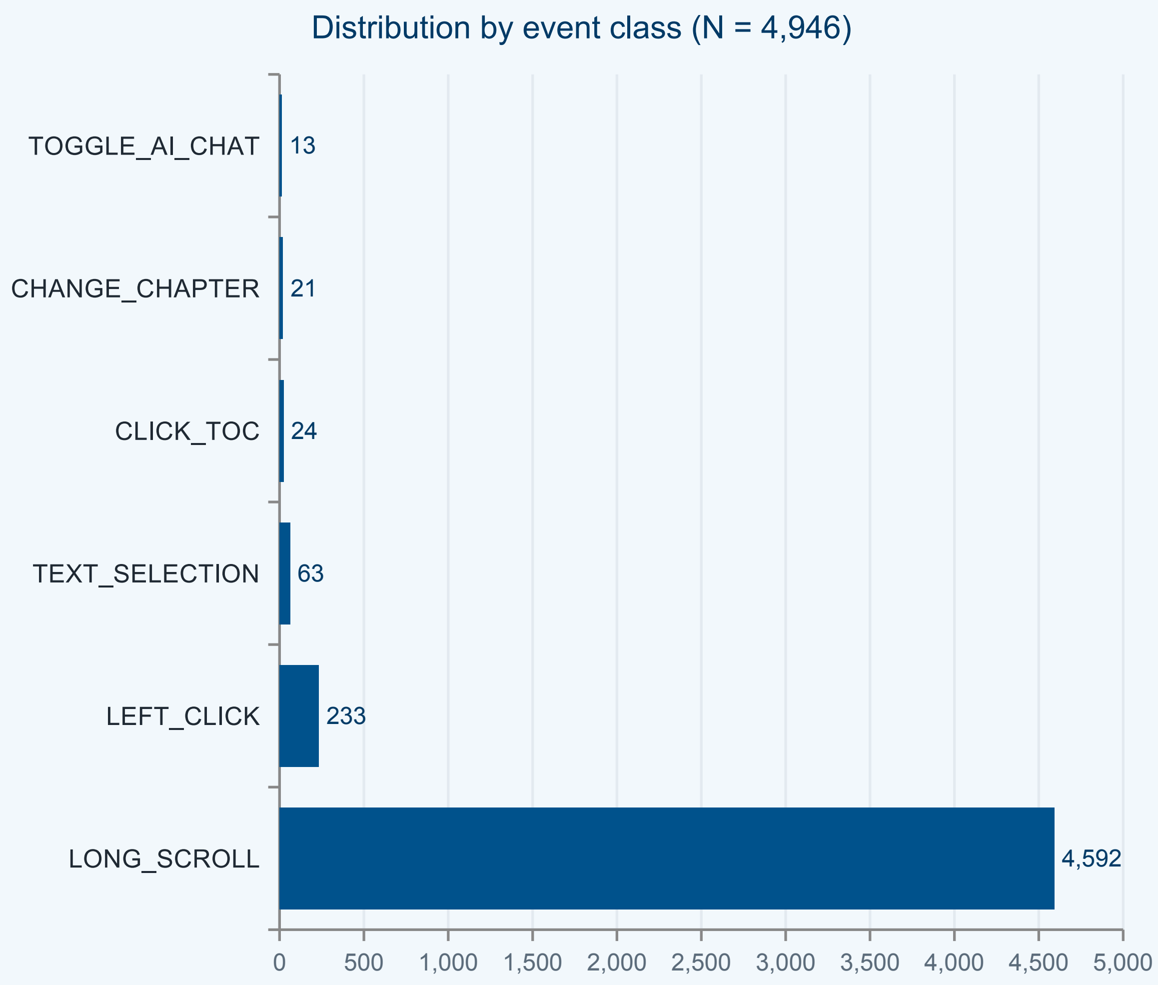
## Five NEU course textbooks running in production

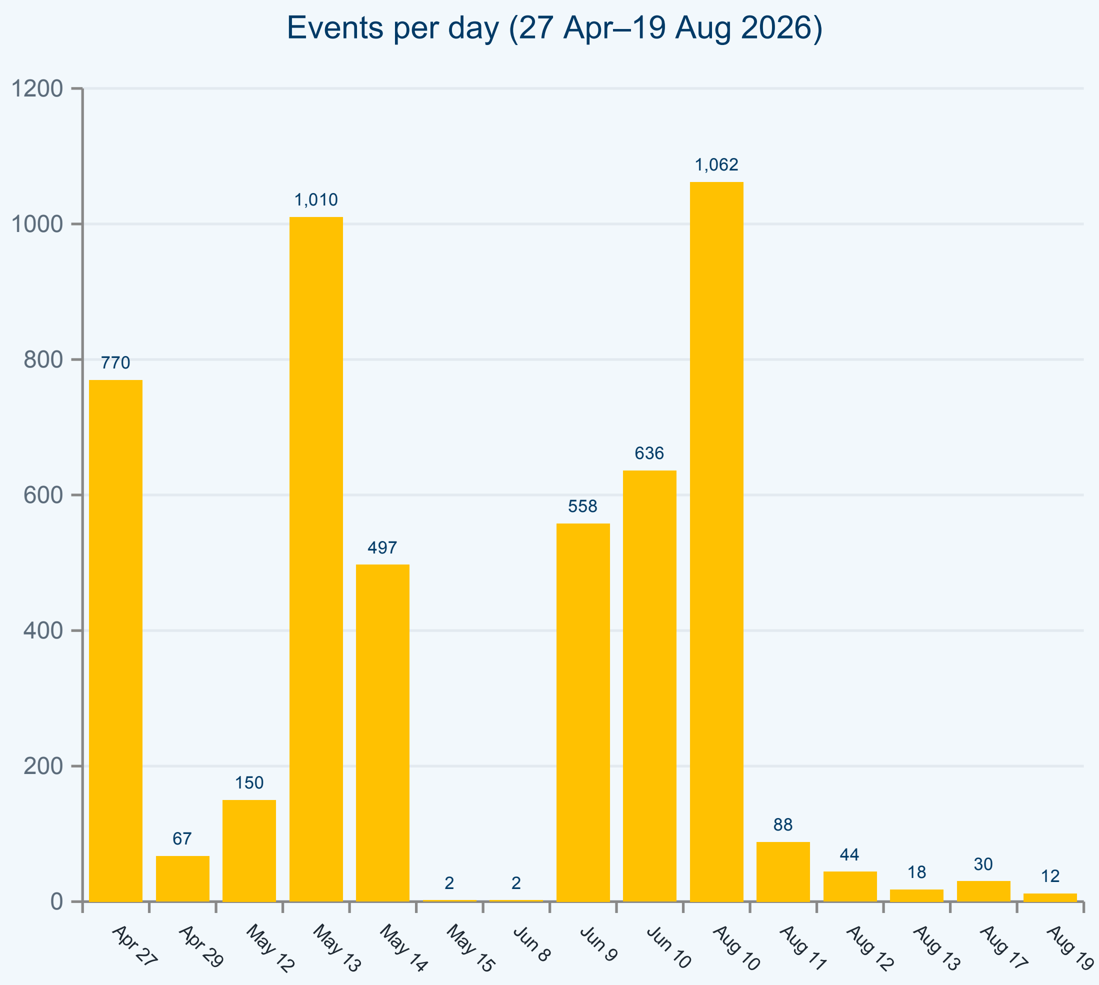
| Textbook | Word pages | Words | Size (MB) |
|---|---|---|---|
| Fundamentals of Programming | 404 | 48,702 | 2.0 |
| Java Programming | 357 | 54,405 | 12.5 |
| Introduction to IT 1 | 324 | 67,687 | 19.7 |
| Web Design | 342 | 69,175 | 53.3 |
| Software Security & Database Practice | 73 | 8,067 | 3.9 |
| **Total** | **1,500** | **248,036** | **91.3** |

*Server: Intel Xeon Silver 4310 · 32 GB RAM · Ubuntu 24.04 LTS · fit.neu.edu.vn/smartdoc*

# 4,946 events over 15 active days





# Key results at a glance



*Infrastructure results — the premise this work sets up; outcome-level results are ongoing.*

### 1 Production-grade, not a testbed

5 textbooks, 1,500 pages, zero manual fixes.

### 2 Real multi-student adoption

30 students, two instructors — real independent use.

### 3 Passive reading, real help-seeking

92.8% long-scroll; 13 AI-tutor opens mid-read.

### 4 Server headroom holds

~5.1 GB after 706h uptime — still ample.

# Pilot feedback and threats to validity



## PILOT FEEDBACK

- Zero-friction ingestion — days of work cut to minutes.
- Students praised the AI tutor for clarifying terms.
- Complaint: pagination wait on low-end mobile devices.
- SUS/TAM user study left to future work.

## THREATS TO VALIDITY

- One faculty, modest student count — limited generalizability.
- Self-hosted Ollama ties risk to available GPU capacity.
- Formatting judged by authors, not blinded external raters.

# Architectural limitations and future work



## LIMITATIONS

- Cold-cache latency — Redis is the natural next step.
- Single on-premises server — a single point of failure.
- Retrieval scope fixed at document level; ablation is future work.

## PEDAGOGICAL FUTURE WORK

- Predict shallow vs. deep reading from scroll + dwell time.
- Mine selections into concept-difficulty heatmaps.
- Trigger adaptive interventions on shallow-reading detection.
- Controlled study with SUS/TAM beyond informal feedback.

# Key sources cited in this work



[1] **Lewis, P., et al.** Retrieval-augmented generation for NLP tasks. NeurIPS 33 (2020)

[2] **Bidelman, E.** Headless Chrome for server-side rendering. Google Chrome Blog (2018)

[3] **Siemens, G., Long, P.** Penetrating the fog: analytics in education. EDUCAUSE Review 46(5) (2011)

[4] **Zawacki-Richter, O., et al.** Systematic review of AI in higher education. IJETHE 16, 39 (2019)

[5] **Jacobs, S., Jäschke, S.** GPT-4 and retrieval-augmented generation. arXiv:2405.06681 (2024)

[6] **Ji, Z., et al.** Survey of hallucination in NLG. ACM Comp. Surveys 55(12) (2023)

[7] **Holmes, W., Bialik, M., Fadel, C.** AI in Education. Center for Curriculum Redesign (2019)

[8] **Bates, A.W.** Teaching in a Digital Age, 2nd edn. BCcampus (2019)

[9] **World Wide Web Consortium** CSS Paged Media Module Level 3. W3C (2023)

# Thank you

*Questions & discussion welcome*

| | |
|---|---|
| **PRESENTED BY** | Nguyễn Đức Thuận |
| **LIVE DEPLOYMENT** | fit.neu.edu.vn/smartdoc |
| **AFFILIATION** | Faculty of Information Technology, NEU |
| **VENUE** | ICETAI 2026 · National Economics University, Hanoi |