Weekly Status Report

WEEK 3

UGP1 | MSE 496

Major Goals

DONE

ONGOING

01

EXTRACTING PAPERS AND COMPLETING THE DATASET

02

AGENTIC ORCHESTRATION

To experiment with agentic orchestration and develop agents for property extraction

03

PHYSICAL STUDY OF ALD

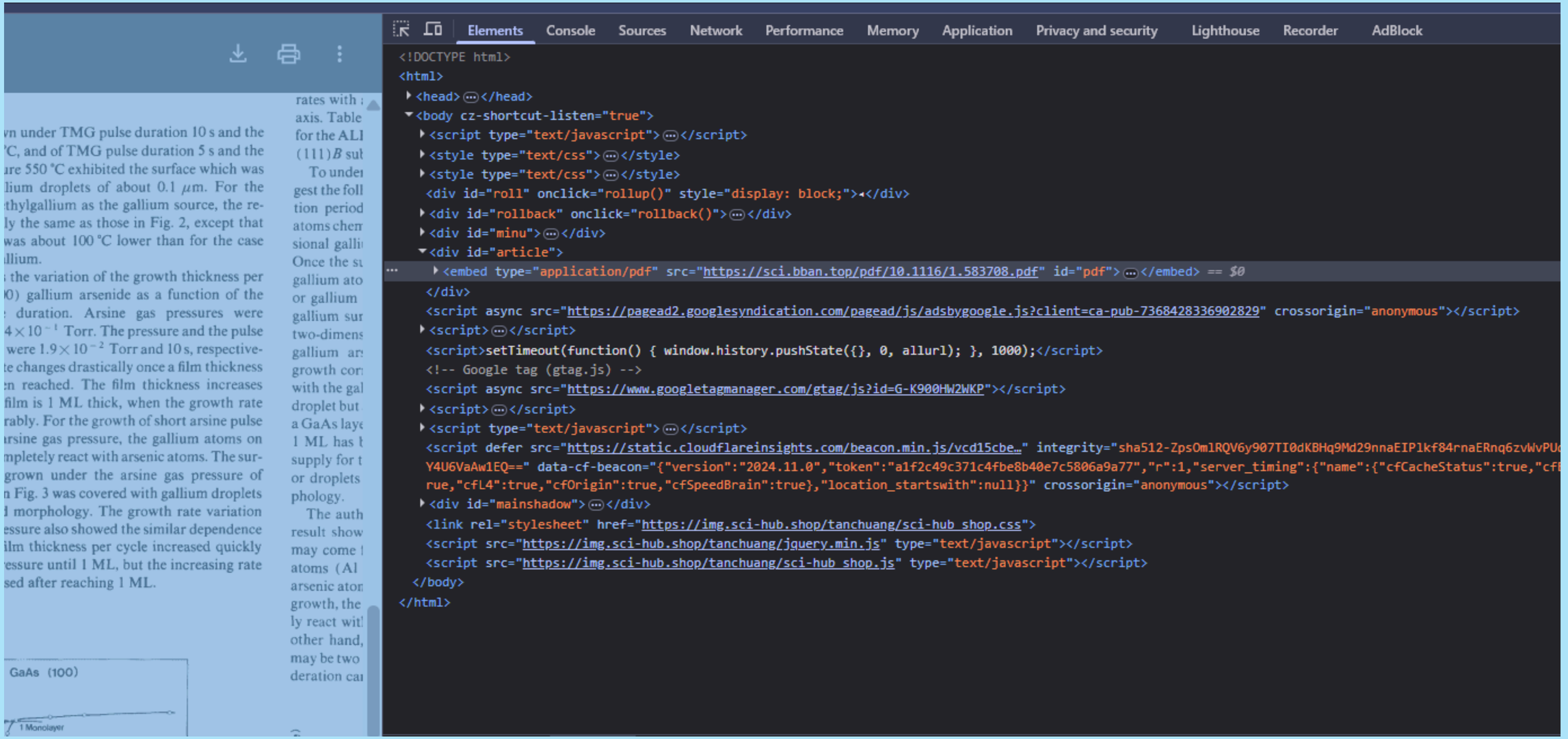
To understand the physics behind Atomic Layer Deposition to build the relevant agents.

1

DOWNLOADING AND SAVING THE PAPERS

GENERAL STATUS

USING SCIHUB TO DOWNLOAD THE PAPERS



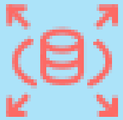
HTML BASED PDF EXTRACTION

DOWNLOAD PHILOSOPHY

```python
for reference in tqdm(references):
    doi = reference.get("reference_doi")
    process_id = reference.get("process_id")
    process = process_map.get(process_id, {})

    try:
        file_name = save2pdf(doi)
        status = "downloaded"
        error = None
    except Exception as e:
        file_name = None
        status = "exception"
        error = str(e)

    if not file_name:
        failed.append({
            "reference_doi": doi,
            "process_id": process_id,
            "status": status,
            "error": error
        })
        time.sleep(random.uniform(1.6, 2.2))
        continue

    successful.append({
        "file_name": file_name,
        "process_id": process_id,
        "reference_doi": doi,
        "process_material": process.get("process_material"),
        "reactantA": process.get("process_reactantA"),
        "reactantB": process.get("process_reactantB"),
        "reactantC": process.get("process_reactantC"),
        "reactantD": process.get("process_reactantD"),
    })

    time.sleep(random.uniform(1.6, 2.2))
```

Extract PDFs from SciHub

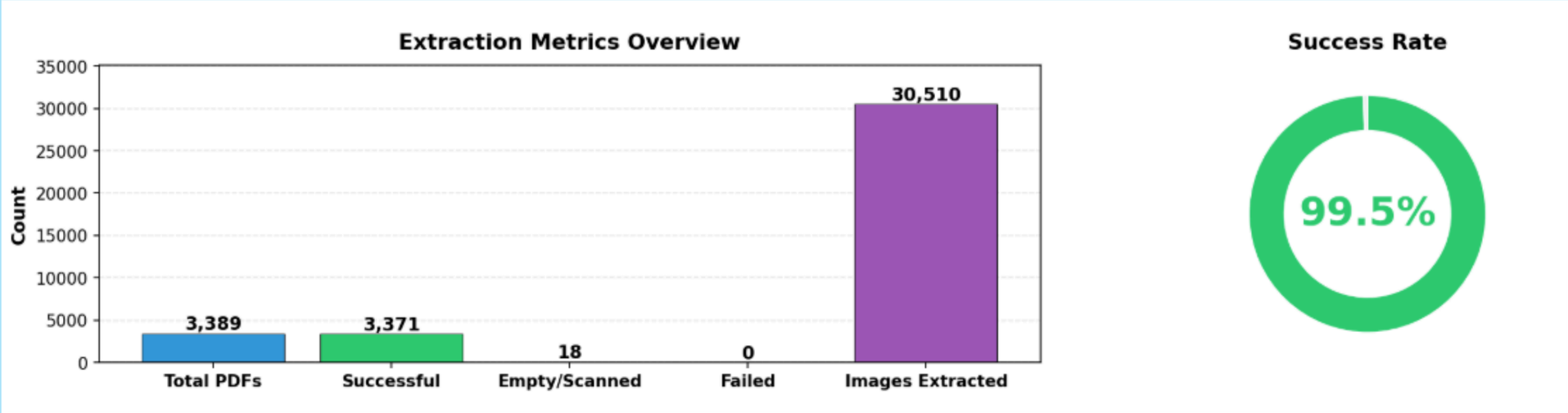
Pipeline Diverges



Save Failed References

Save Successful Metadata

EXTRACTION VISUALISATIONS



TOTAL OF 3,389 PDFS WERE EXTRACTED

SUCCESSFUL METADATA

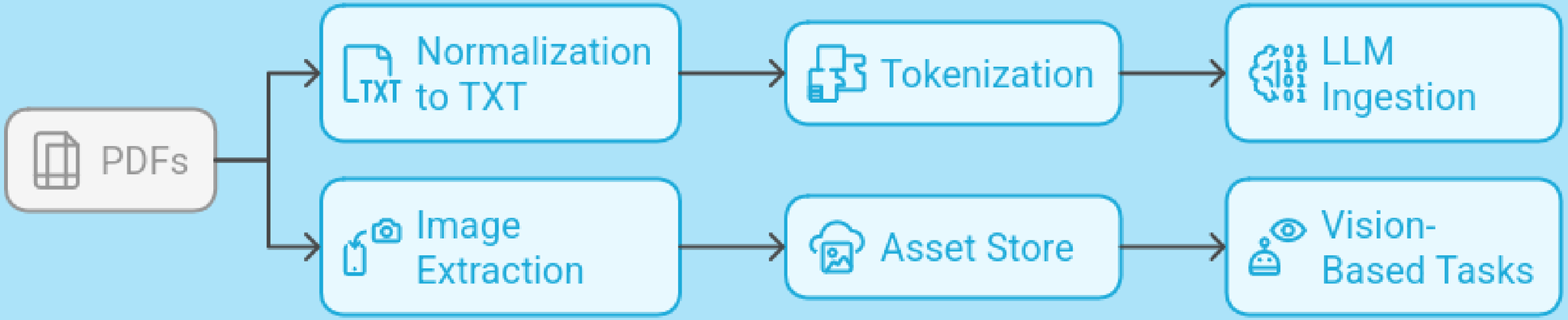
	A	B	C	D	E	F	G	H
	file_name	process_id	reference_doi	process_materia	reactantA	reactantB	reactantC	reactantD
	10.1002_1521-3	429	10.1002/1521-3	ZrO2	Zr(OtBu)4	H2O		
	10.1002_(SICI)1	461	10.1002/(SICI)1	Nb2O5	Nb(OEt)5	H2O		
	10.1002_(SICI)1	549	10.1002/(SICI)1	BaS	Ba(thd)2	H2S		
	10.1002_(SICI)1	194	10.1002/(SICI)1	TiN	TiCl4	NH3 + catalyst		
	10.1002_1521-3	348	10.1002/1521-3	GaN	Ga	N2		
	10.1002_1521-3	383	10.1002/1521-3	SrO	Sr(CpiPr3)2	H2O		
	10.1002_1521-3	173	10.1002/1521-3	TiO2	Ti(OiPr)4	H2O		
	10.1002_1521-3	193	10.1002/1521-3	TiN	TiCl4	NH3		
	10.1002_1521-3	405	10.1002/1521-3	Y2O3	Y(thd)3	O3		
	10.1002_1521-3	414	10.1002/1521-3	ZrO2	ZrI4	H2O2		
	10.1109_IITC.20	2240	10.1109/IITC.20	WNxCy	WF6	NH3	BEt3	
	10.1002_zaac.1	220	10.1002/zaac.19	VOx	VOCl3	H2O		
	10.1002_zaac.1	220	10.1002/zaac.19	VOx	VOCl3	H2O		
	10.1016_0040-6	340	10.1016/0040-6	ZnTe	Zn	Te		
	10.1063_1.9227	324	10.1063/1.9227	ZnS	ZnCl2	H2S		
	10.1063_1.9227	674	10.1063/1.9227	Ta2O5	TaCl5	H2O		
	10.1063_1.3317	506	10.1063/1.3317	CdTe	Cd	Te		
	10.1016_0022-0	506	10.1016/0022-0	CdTe	Cd	Te		
	10.1016_0022-0	235	10.1016/0022-0	MnTe	Mn	Te		
	10.1116_1.5723	506	10.1116/1.57235	CdTe	Cd	Te		
	10.1016_0022-0	506	10.1016/0022-0	CdTe	Cd	Te		
	10.1063_1.9534	506	10.1063/1.9534	CdTe	Cd	Te		
	10.1063_1.9534	235	10.1063/1.9534	MnTe	Mn	Te		
	10.1007_BF013	143	10.1007/BF013	POx	POCl3	H2O		
	10.1007_BF013	220	10.1007/BF013	VOx	VOCl3	H2O		
	10.1007_BF013	228	10.1007/BF013	CrOx	CrO2Cl2	H2O		
	10.1007_BF013	96	10.1007/BF013	SiO2	SiCl4	H2O		
	10.1149_1.2114	363	10.1149/1.2114	GaAs	GaMe3	AsH3		
	10.1016_0040-6	321	10.1016/0040-6	ZnO	Zn(OAc)2	H2O		
	10.1016_0040-6	329	10.1016/0040-6	ZnS	Zn(OAc)2	H2S		
	10.1016_0040-6	329	10.1016/0040-6	ZnS	Zn(OAc)2	H2S		

FAILED METADATA

	A	B	C	D
	reference_doi	process_id	status	error
	10.3938/jkps.45.	429	failed	
	10.3938/jkps.34.	520	failed	
	10.3938/jkps.47.	601	failed	
	10.3938/jkps.46.	601	failed	
	10.3938/jkps.42.	601	failed	
	10.3938/jkps.46.	601	failed	
	10.3938/jkps.35.	674	failed	
	10.3938/jkps.45.	712	failed	
	10.3938/jkps.48.	40	failed	
	10.3938/jkps.45.	119	failed	
	10.3938/jkps.46.	536	failed	
	10.3938/jkps.49.	561	failed	
	10.3938/jkps.45.	709	failed	
	10.3938/jkps.33.	19	failed	
	10.3938/jkps.35.	19	failed	
	10.3938/jkps.47.	25	failed	
	10.3938/jkps.46.	25	failed	
	10.3938/jkps.49.	26	failed	
	10.3938/jkps.42.	29	failed	
	10.3938/jkps.48.	29	failed	
	10.3938/jkps.35.	118	failed	
	10.3938/jkps.45.	118	failed	
	10.3938/jkps.47.	119	failed	
	10.3938/jkps.45.	173	failed	
	10.3938/jkps.35.	193	failed	
	10.3938/jkps.49.	310	failed	
	10.3938/jkps.28.	336	failed	
	10.1088/0031-8	235	failed	
	10.1088/0031-8	506	failed	
	10.12693/APhys	2354	failed	
	10.1002/(SICI)1	25	failed	

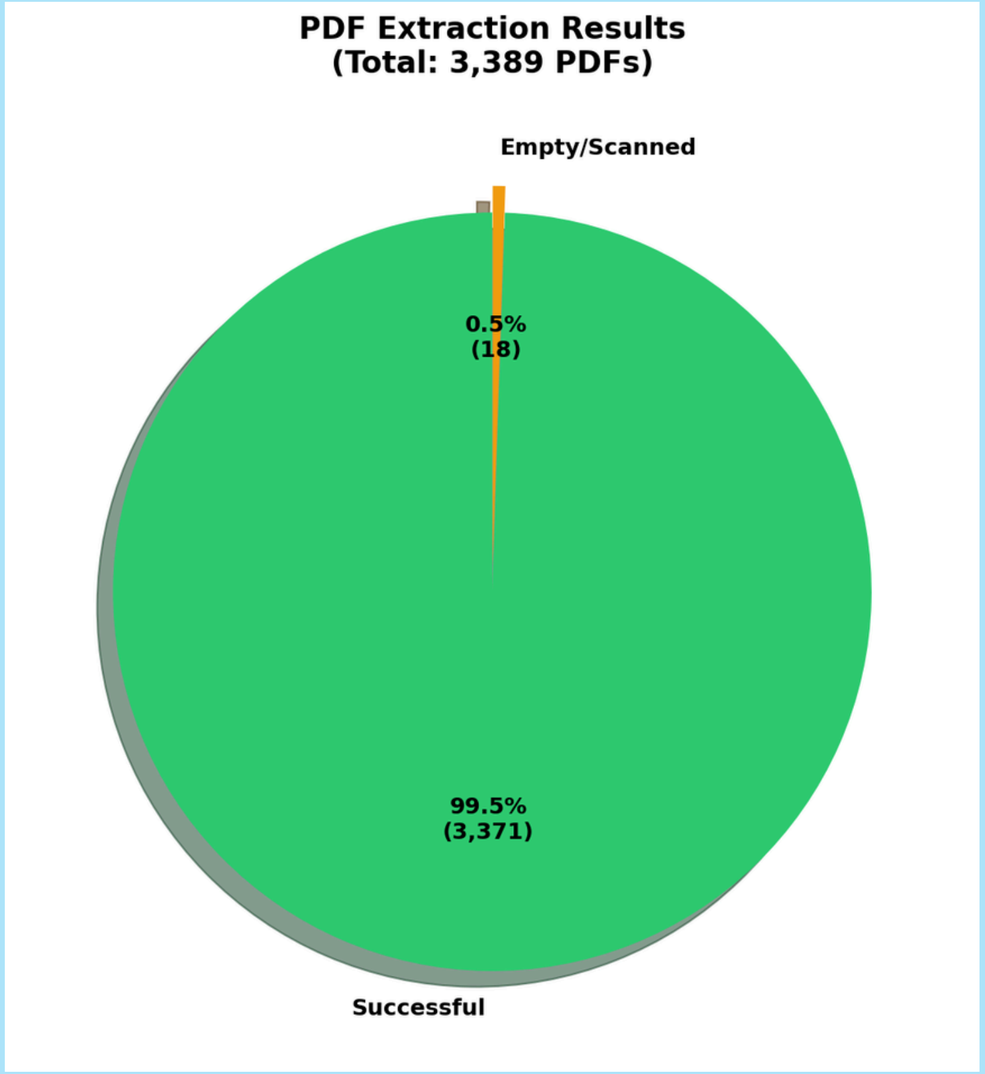
SAVED FOR FUTURE REFERENCE

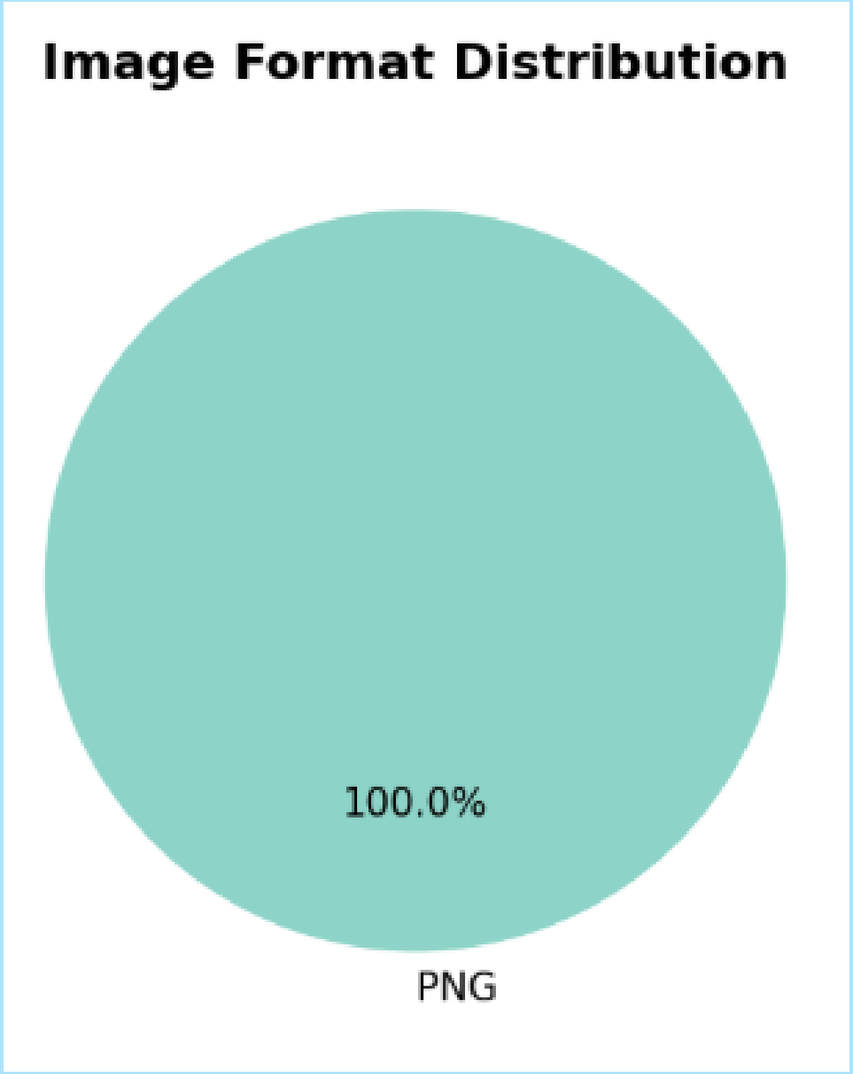
THE PROBLEM: PDF FILES



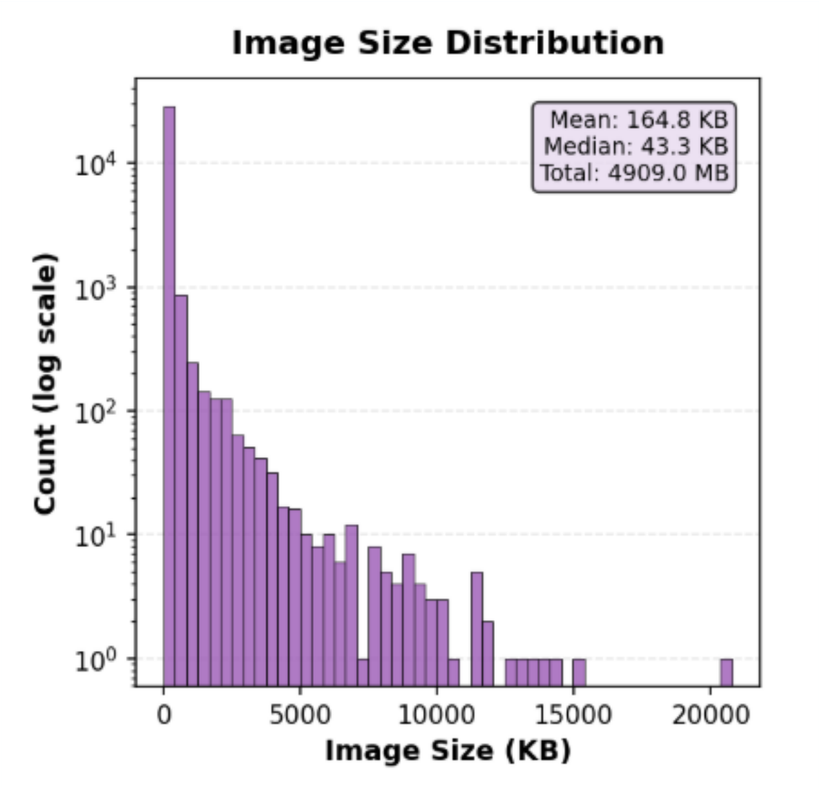
PDF FILES ARE UNSTRUCTURED AND TOUGH TO USE FOR DOWNSTREAM TASKS.

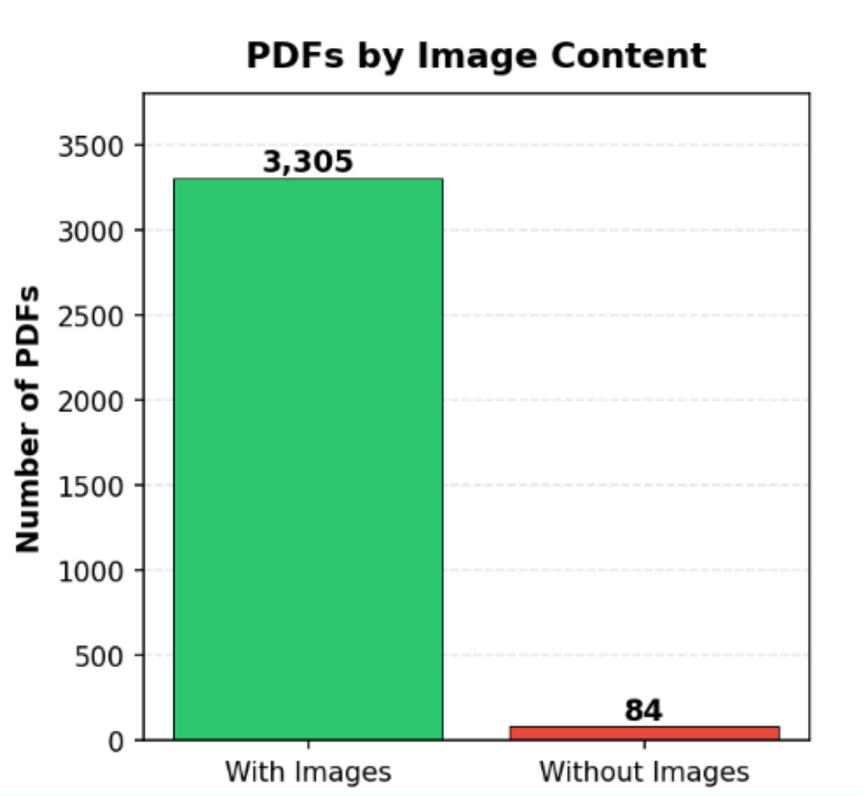
CONVERSION STATISTICS



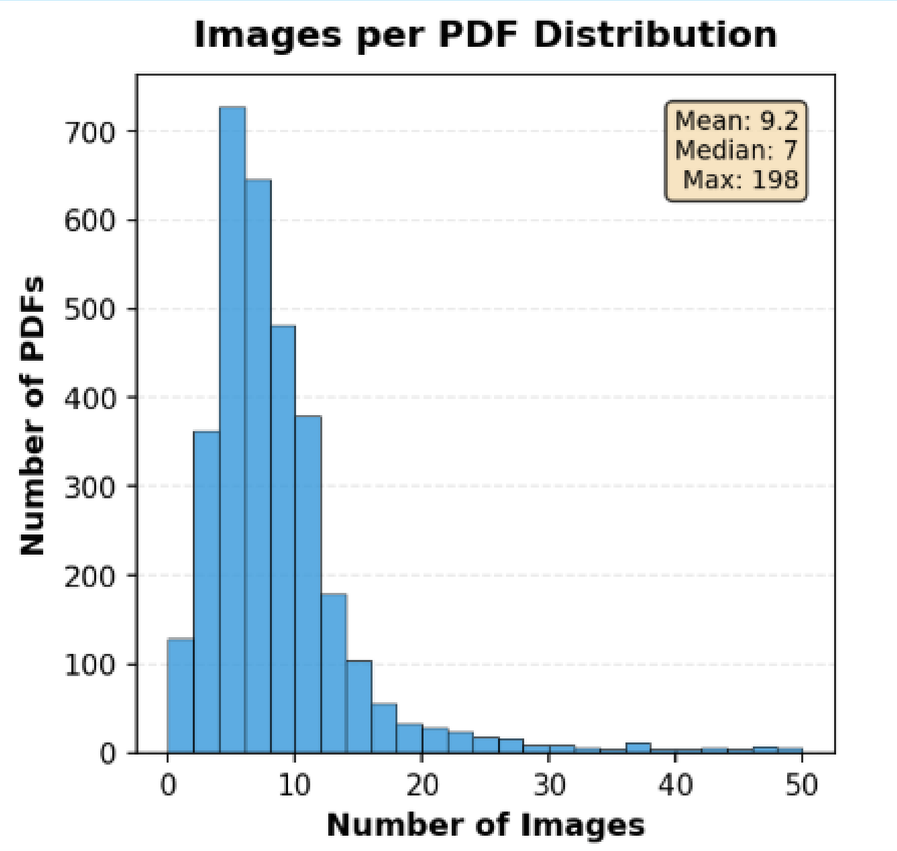


CONVERSION STATISTICS





CONVERSION STATISTICS



Extraction Density

9.1

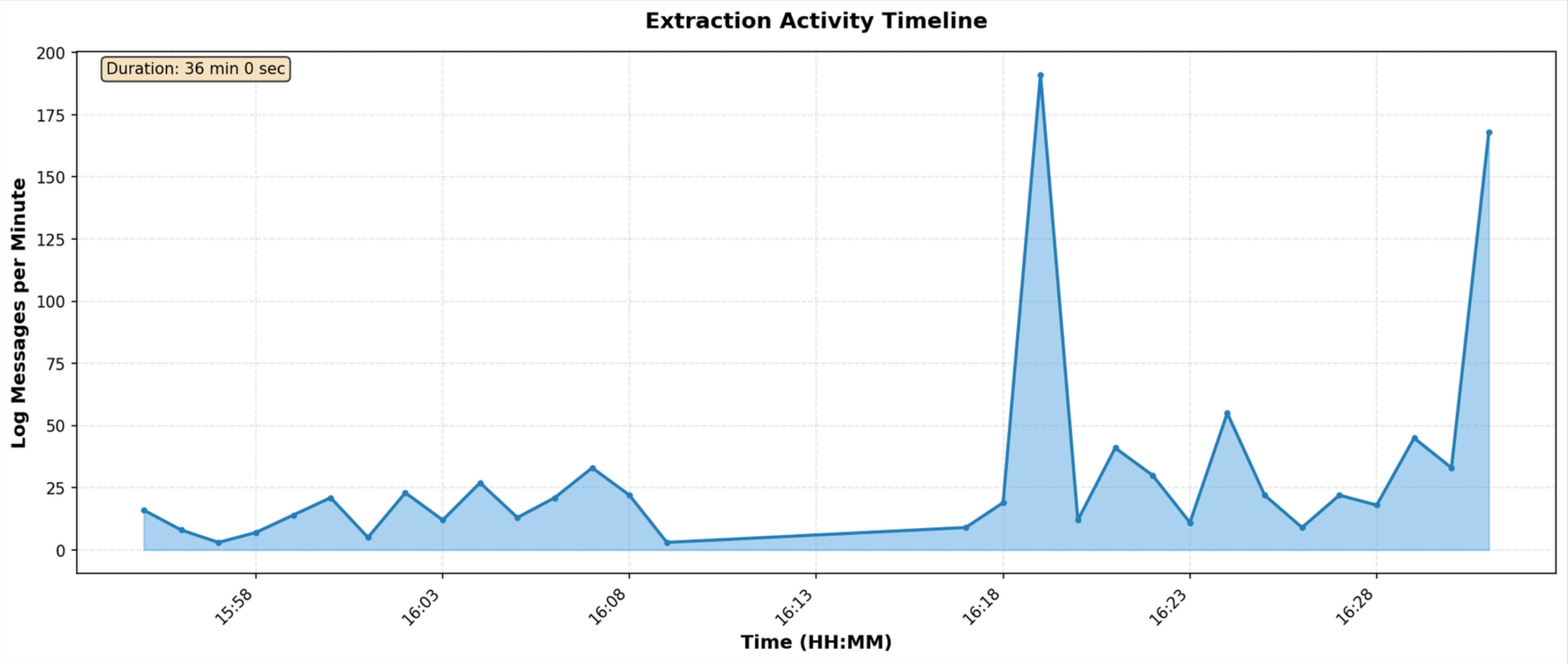
Images per PDF
(average)

Empty/Scanned PDFs:

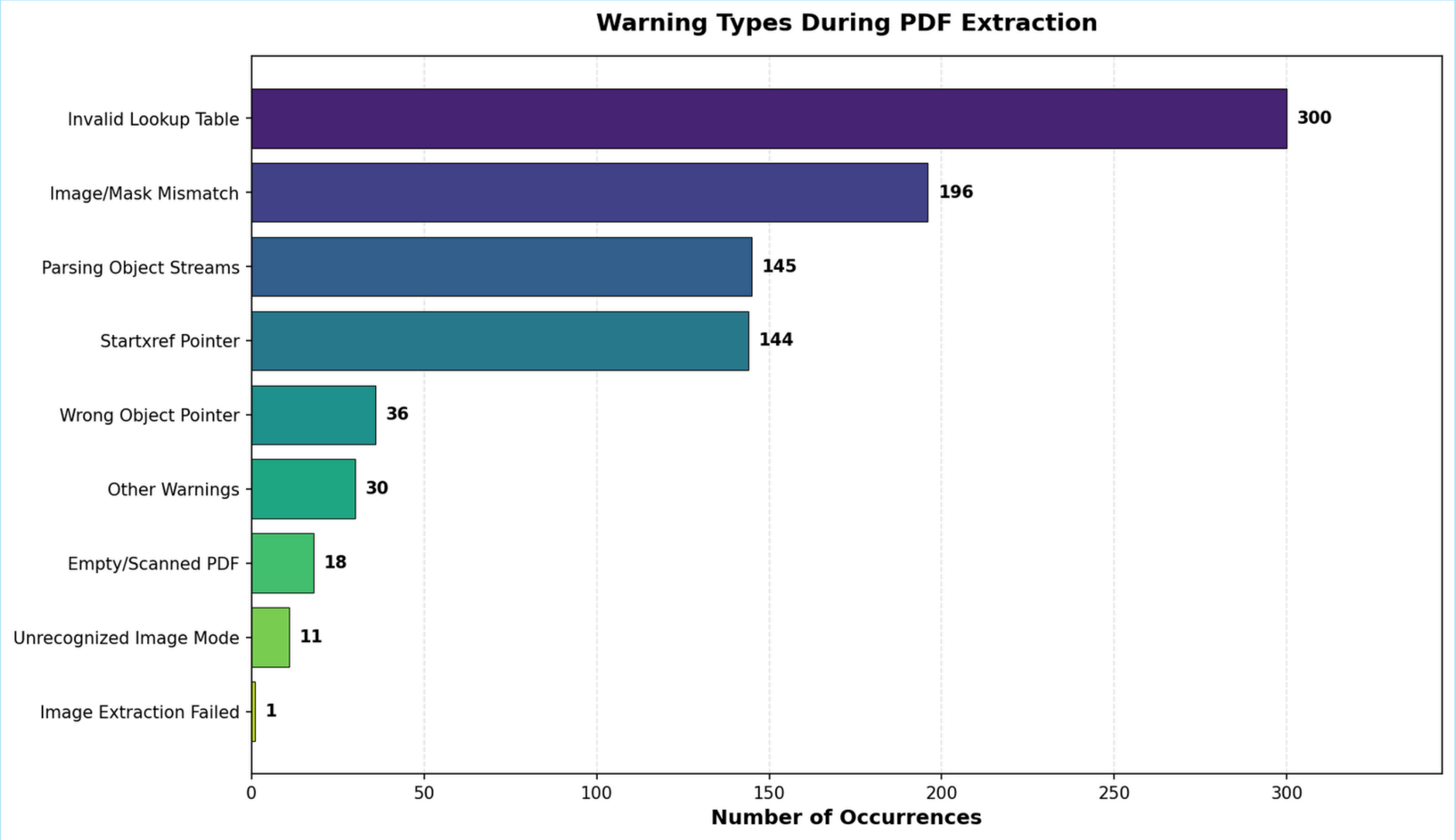
- paper606.pdf
- paper616.pdf
- paper605.pdf
- paper617.pdf
- paper619.pdf
- paper612.pdf
- paper615.pdf
- paper614.pdf
- paper610.pdf
- paper618.pdf
- paper611.pdf
- paper608.pdf
- paper613.pdf
- paper607.pdf
- paper609.pdf

... and 3 more

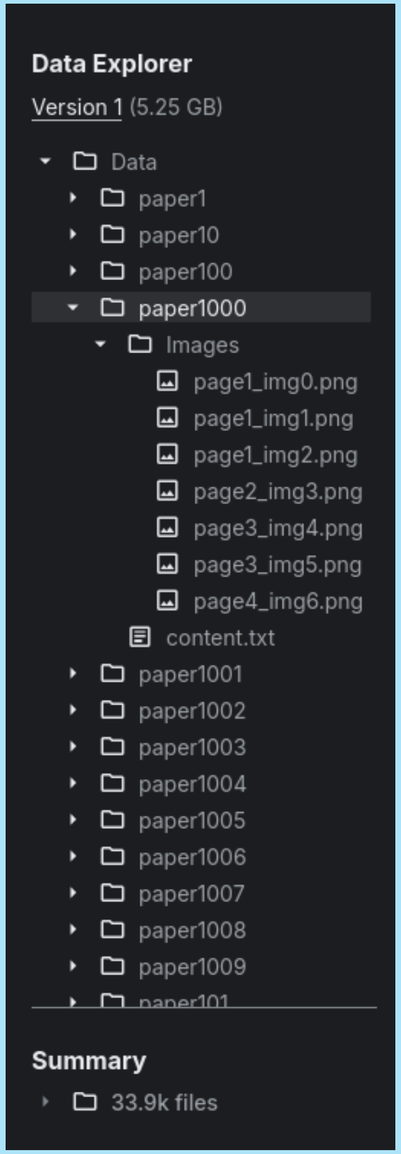
EXTRACTION TIMELINE



ERROR ANALYSIS



FINAL DATASET FORMAT UPLOADED TO KAGGLE



HTTPS://WWW.KAGGLE.COM/DATASETS/DABEET/EXTRACTED-ATOMIC-LAYER-DEPOSITION

EASE OF PROCESSING

[4]:

```
with open("/kaggle/input/datasets/dabeet/extracted-atomic-layer-deposition/Data/paper1/content.txt",'r') as f:
    content = f.read()
    print(content)
```

```
Communications
Low Temperature CVD of Crystalline Titanium
Dioxide Films Using Tetranitratotitanium(IV)**
By David C. Gilmer, Daniel G. Colombo,
Charles J. Taylor, Jeff Roberts, Greg Haugstad,
Stephen A. Campbell, Hyeon-Seag Kim, Glen D. Wilk,
Michael A. Gribelyuk, and Wayne L. Gladfelter*
The continuing push to decrease the size of microelectronic devices is hampered by some of the physical properties of the current materials. Sil
icon dioxide is currently
used as the gate dielectric in metal oxide semiconductor
field effect transistors (MOSFETs), and operation of this
device requires that the thickness of the dielectric be scaled
along with the length of the gate between the source and
drain. As the gate lengths approach 0.1 mm, the required
SiO
2 thickness will drop to 15–20 Å, and leakage current
through the dielectric will rise to unacceptable limits.
One alternative is to replace SiO 2 (dielectric constant k =
3.9) with a material having a higher dielectric constant that
will allow the use of thicker, less leaky, films. Towards this
end compounds such as Ta
2O5 (20 £k £ 25) have been evaluated. We have recently reported on the first successful
metal insulating semiconductor field effect transistor (MISFET) using TiO
2 (25 £ k £ 30) grown from Ti(O- iPr)4
(tetraisopropoxotitanium(IV) as the gate dielectric.
Titanium dioxide has been deposited from single-source
precursors, such as Ti(OR)4, or from TiCl4, where the latter
requires a source of oxygen (i.e., O2 or H2O). With few
exceptions, depositions at temperatures below approximately 250 °C lead to amorphous films, which typically exhibit lower dielectric constants.
Depending on the source
and conditions, carbon or chlorine impurities can be present in the films. In addition, the presence of O 2
and/or H2O in the reactor can lead to unwanted oxidation
of the silicon wafer before TiO 2 is formed. To surmount
these problems it was desirable to have a precursor that
will deposit pure, crystalline TiO2 at low temperature without requiring a separate source of oxygen.
```

2

AGENTIC ORCHESTRATION

GENERAL STATUS

AGENT 1: MATERIAL EXTRACTOR

```python
from langchain_core.prompts import PromptTemplate

prompt = PromptTemplate.from_template("""

You are an expert Materials Science Data Extraction Assistant.
Your task is to extract all specific material entities from the provided text.

Target Entities:
1. Chemical Formulas: (e.g., Bi2Te3, SiO2, CH3NH3PbI3)
2. Doped/Alloyed Systems: (e.g., SnSe:Na, PbTe–AgSbTe2, In_xGa_(1-x)As)
3. Specific Material Names: (e.g., Graphene, Borophene, Black Phosphorus, Gold, Silicon)
4. Polymers & Acronyms: (e.g., PEDOT:PSS, PMMA, PTFE, CNTs)
5. Specific Phases/Crystal Structures: (e.g., α-Fe2O3, 1T'-MoTe2, cubic-SiC)

Extraction Rules:
- Exactness: Preserve the string exactly as written, including capitalization, hyphens, and subscripts (e.g., keep "n-type Si" or "1T-TaS2").
- Completeness: Include dopants and stoichiometric variables if they are part of the name (e.g., capture "Al-doped ZnO" rather than just "ZnO")
- Exclusions: - Ignore generic categories (e.g., "semiconductor", "metal", "alloy", "ceramic").
    - Ignore device components unless chemically specified (e.g., ignore "substrate", "anode", "gate", but keep "SiO2 substrate" -> extract "Si
    - Ignore units of measurement or concentrations alone (e.g., "10 nm", "5%").
- Deduplication: Remove exact duplicates.
- Limit: Return only the first {max_materials} distinct materials found.

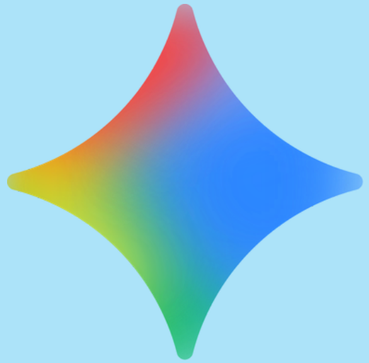
Output Format:
Return valid JSON only. No markdown formatting, no conversational text.

{{
    "materials": ["Material_String_1", "Material_String_2", ...]
}}

Text to Analyze:
{fulltext}

""")

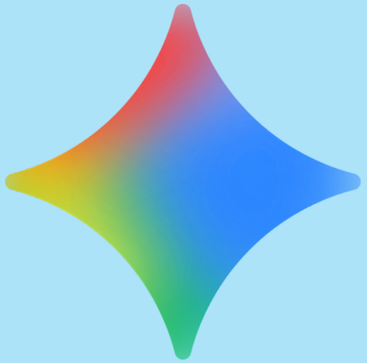
out = model.invoke(prompt.format(fulltext=content, max_materials=700))
```



GEMINI AS A MATERIAL EXTRACTOR

AGENT 1: MATERIAL EXTRACTOR


```
{
    "materials": [
        "Titanium Dioxide",
        "Tetranitratotitanium(IV)",
        "Silicon dioxide",
        "SiO 2",
        "Ta 2O5",
        "TiO 2",
        "Ti(O- iPr)4",
        "tetraisopropoxotitanium(IV)",
        "Ti(OR)4",
        "TiCl4",
        "O2",
        "H2O",
        "carbon",
        "chlorine",
        "silicon",
        "TNT",
        "p-Si(100)",
        "Ar",
        "TiO 2.0–0.1",
        "nitrogen",
        "anatase phase",
        "anatase",
        "Ti(OiPr)4",
        "Ti(NO3)4",
        "silicon oxide",
        "titanium oxide",
        "Si(100)",
        "Pt/TiO 2/p-Si(100)/Al",
        "Pt",
        "Al",
        "p-type Si",
        "oxygen",
        "H 2",
        "H2",
        "Ti(NO 3)4",
        "Poly[bis((methoxyethoxy)ethoxy)-phosphazene]",
        "MEEP",
        "lithium",
        "platinum",
        "aluminum",
        "Ti(O-iPr)4",
        "O 2"
    ]
```

GEMINI AS A MATERIAL EXTRACTOR

AGENT 2: PAPER SUMMARISER

```
from langchain_core.prompts import PromptTemplate

prompt = PromptTemplate.from_template("""
You are an expert Atomic Layer Deposition (ALD) Scientific Summarisation Assistant.
Your task is to provide a high-level synthesis of the provided research paper, focusing on the scientific narrative rather than specific proces

Summarisation Objectives:

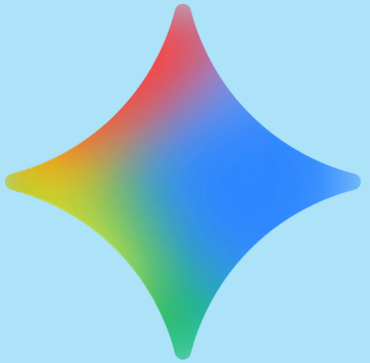
1. Executive Summary: Provide a concise overview of the study's objective, the material system investigated, and the primary breakthrough.
2. Logical Flow of the Paper: Explain the structural progression of the research. How does the paper move from the initial hypothesis or probler
3. Core Methodology & Logic: Summarize the *approach* taken (e.g., comparing two different precursors or investigating the effect of plasma powe
4. Key Findings: What are the most significant conclusions regarding film quality, interface physics, or material performance?
5. Impact & Application: Why does this work matter to the ALD community or the target industry?

Summarisation Rules:
- Avoid Granularity: Do NOT list specific growth per cycle (GPC) values, exact temperatures, or pulse times unless they are the central discover
- Scientific Precision: Maintain formal nomenclature and exact chemical formulas.
- Structural Clarity: Focus on the "Why" and "How" rather than just the "What."
- No Hallucinations: Summarize only what is explicitly stated.

Output Format (JSON only):
{{
  "executive_summary": "High-level overview of the research.",
  "paper_flow": "A step-by-step explanation of the paper's logical structure and progression.",
  "methodological_approach": "The general experimental strategy used by the authors.",
  "key_findings": "The most important scientific results or physical insights discovered.",
  "significance_and_applications": "The broader impact and intended use cases for this research."
}}

Text to Analyze:
{fulltext}
""")

out = model.invoke(prompt.format(fulltext=content, max_materials=700))
```

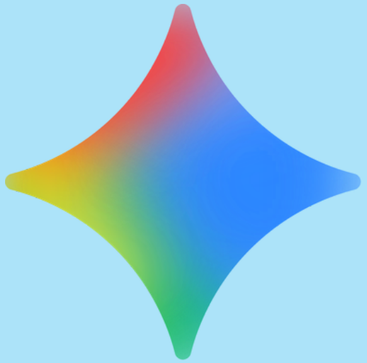


GEMINI AS A PAPER SUMMARISER

AGENT 2: PAPER SUMMARISER

```
print(out.content[0]["text"])
```

```
```json
{
  "executive_summary": "The study investigates the use of tetranitratotitanium(IV) [Ti(NO3)4] as a novel single-source precursor for the chemical vapor deposition (CVD) of titanium dioxide (TiO2) thin films. The primary breakthrough is the synthesis of highly pure, crystalline anatase-phase TiO2 at exceptionally low temperatures—as low as 184 °C—without the need for external oxidizing agents like water or oxygen. This approach addresses the limitations of conventional precursors that typically require high thermal budgets or lead to unwanted interfacial oxidation of silicon substrates.",
  "paper_flow": "The research follows a logical progression from problem identification to device-level validation. It begins by establishing the necessity of high-k dielectrics to replace SiO2 in scaled MOSFET devices. The authors then justify the selection of Ti(NO3)4 based on its volatility and anhydrous nature. The narrative proceeds to demonstrate thin film growth using both LPCVD and UHV-CVD techniques, followed by a rigorous characterization of crystallinity, purity, and morphology. The paper concludes by fabricating metal-insulator-semiconductor (MIS) capacitors to evaluate the dielectric performance and the impact of post-deposition annealing on leakage current.",
  "methodological_approach": "The experimental strategy centers on evaluating a single-source precursor that contains its own oxygen source within the molecular structure, thereby eliminating the competition between film growth and substrate oxidation by external oxidants. The researchers compared films grown across different pressure regimes (LPCVD vs. UHV-CVD) to determine the influence of deposition environment on growth rates and film quality. Furthermore, they utilized a comparative analysis, benchmarking the results against films grown from standard precursors like tetraisopropoxotitanium to highlight the superior grain size and smoothness achieved at lower temperatures.",
  "key_findings": "The study reveals that Ti(NO3)4 enables the growth of stoichiometric, carbon-free, and nitrogen-free TiO2 films. A significant physical insight is the achievement of the crystalline anatase phase at temperatures where TiO2 is typically amorphous. Morphologically, the films exhibit larger grain diameters and lower RMS roughness than those derived from alkoxide precursors. Electrically, while an unavoidable interfacial oxide layer forms due to the precursor's oxidizing nature, the films demonstrate high dielectric constants. Notably, a dual-step annealing process (oxygen followed by hydrogen) was found to reduce leakage current density by six orders of magnitude.",
  "significance_and_applications": "This work is highly significant for the semiconductor industry as it provides a pathway for integrating high-k TiO2 dielectrics into devices with strict thermal budgets. By enabling low-temperature crystallinity and high purity, the research facilitates the development of high-performance MISFETs. The use of anhydrous metal nitrates like TNT offers a broader chemical strategy for depositing other metal oxides where maintaining a low-temperature profile and avoiding external water/oxygen sources are critical requirements."
}
```
```



GEMINI AS A PAPER SUMMARISER

3

PHYSICAL STUDY OF ATOMIC LAYER DEPOSITION

GENERAL STATUS

REFERENCE TEXT

Materials Today • Volume 17, Number 5 • June 2014 RESEARCH

ELSEVIER

CrossMark

RESEARCH: Review

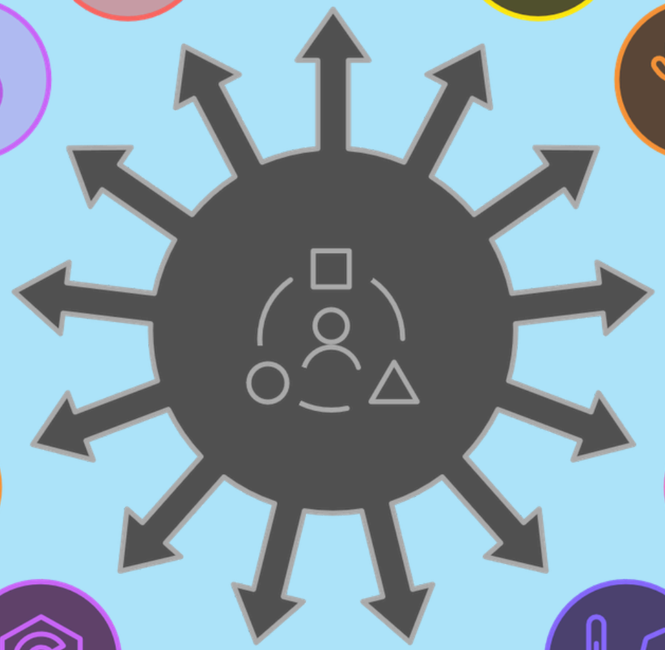
A brief review of atomic layer deposition: from fundamentals to applications

Richard W. Johnson[1,3], Adam Hultqvist[2,3] and Stacey F. Bent[1,2,*]

[1] Department of Materials Science and Engineering, Stanford University, Stanford, CA 94305, USA
[2] Department of Chemical Engineering, Stanford University, Stanford, CA 94305, USA

Atomic layer deposition (ALD) is a vapor phase technique capable of producing thin films of a variety of materials. Based on sequential, self-limiting reactions, ALD offers exceptional conformality on high-aspect ratio structures, thickness control at the Angstrom level, and tunable film composition. With these advantages, ALD has emerged as a powerful tool for many industrial and research applications. In this review, we provide a brief introduction to ALD and highlight select applications, including $Cu(In,Ga)Se_2$ solar cell devices, high-k transistors, and solid oxide fuel cells. These examples are chosen to illustrate the variety of technologies that are impacted by ALD, the range of materials that ALD can deposit – from metal oxides such as $Zn_{1-x}Sn_xO_y$, ZrO_2, Y_2O_3, to noble metals such as Pt – and the way in which the unique features of ALD can enable new levels of performance and deeper fundamental understanding to be achieved.

ALD PARAMETERS



Precursor and Coreactant
The chemical building blocks for film growth

Temperature and Pressure
Environmental conditions influencing reaction rates

Reaction Time
Duration of each deposition cycle

Purging Agents
Inert gases to remove residual reactants

Repetition and Kinetics
Number of cycles and reaction speed

ALD Temperature Window
Optimal temperature range for film growth

Growth per Cycle
Thickness of film deposited in each cycle

Deposition Rate
Speed of film deposition

Composition Control
Ensuring desired film composition

Pathway Restrictions
Limitations due to available reaction routes

Precursor Types
Metal-organic vs. Inorganic precursors

No Universal Best
Lack of a single optimal precursor system

Tradeoffs
Balancing various factors for optimal results

Thank you