

Researcher Assistant Service

- 연구자를 위한 통합 연구 플랫폼 -

데이터 엔지니어링 1조 김송일 박현민 박지원

INDEX

#01 서론

#1-01 프로젝트 배경

#1-02 프로젝트 목표

#02 기능 소개

#2-01 Mailing

#2-02 개인 대시보드

#03 기술

#3-01 DB 구성

#3-02 Data & Technology

#3-03 Mailing Airflow

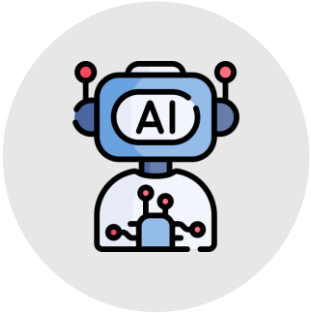
#3-04 Report Generation

#04 결론

#4-01 의의 및 확장방안

#05 Demo

#1-01 프로젝트 배경



AI 기반 연구 활성화에 따른
논문 출판 속도 가속화



연구자별 맞춤형
논문 선별, 요약 도구 부재



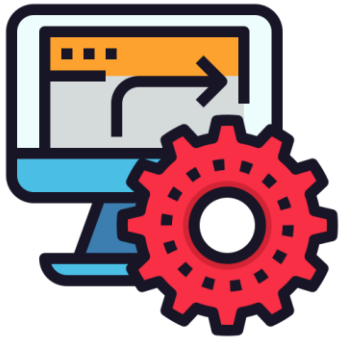
연구 생산성을 향상을 위해
존재하는 여러 연구 보조 도구

1) 최신 연구 동향 파악과 2)연구 라이프사이클 전반 관리를 가능하게 하는 서비스가 있으면 어떨까?

#1-02 프로젝트 목표

연구자가 관련 연구 동향 파악부터 검색, 논문 정리, 프로젝트 일정 관리까지
연구 전체 라이프 사이클을 하나의 플랫폼에서 관리하는 연구자 지원 서비스 개발

자동화 (Automation)



매일/매주 최신 논문 트렌드를
분석하여 이메일로 자동 전송

통합 (Integration)



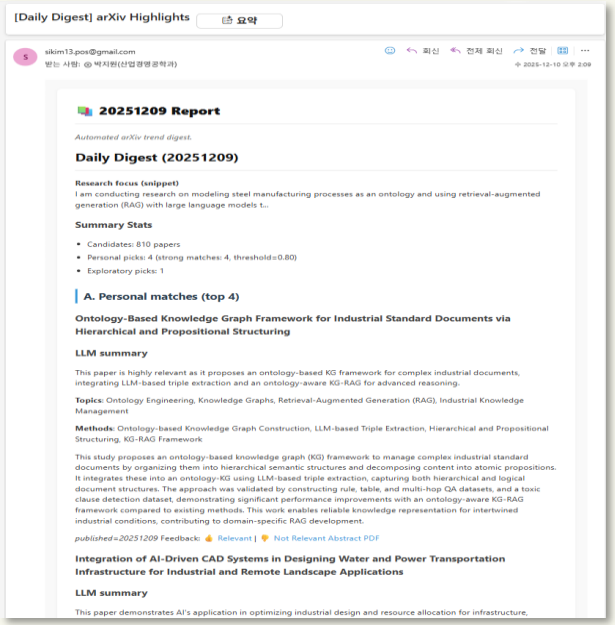
논문 검색-정리-일정-기록을
하나의 플랫폼에서 통합하여
연구 흐름이 끊기지 않도록 지원

데이터화 (Data Storage)



연구 활동 데이터를 기록하여
연구 지식과 개인 데이터를 자산으로 전환

#2-01 Mailing - Daily Digest


[Daily Digest] arXiv Highlights 요약

sikim13.pos@gmail.com
받는 사람: 박지원(산업경영공학과)

회신 전체 회신 전달
수 2025-12-10 오후 2:09

20251209 Report

Automated arXiv trend digest.

Daily Digest (20251209)

Research focus (snippet)
I am conducting research on modeling steel manufacturing processes as an ontology and using retrieval-augmented generation (RAG) with large language models t...

Summary Stats

- Candidates: 810 papers
- Personal picks: 4 (strong matches: 4, threshold=0.80)
- Exploratory picks: 1

A. Personal matches (top 4)

Ontology-Based Knowledge Graph Framework for Industrial Standard Documents via Hierarchical and Propositional Structuring

LLM summary

This paper is highly relevant as it proposes an ontology-based KG framework for complex industrial documents, integrating LLM-based triple extraction and an ontology-aware KG-RAG for advanced reasoning.

Topics: Ontology Engineering, Knowledge Graphs, Retrieval-Augmented Generation (RAG), Industrial Knowledge Management

Methods: Ontology-based Knowledge Graph Construction, LLM-based Triple Extraction, Hierarchical and Propositional Structuring, KG-RAG Framework

This study proposes an ontology-based knowledge graph (KG) framework to manage complex industrial standard documents by organizing them into hierarchical semantic structures and decomposing content into atomic propositions. It integrates these into an ontology-KG using LLM-based triple extraction, capturing both hierarchical and logical document structures. The approach was validated by constructing rule, table, and multi-hop QA datasets, and a toxic clause detection dataset, demonstrating significant performance improvements with an ontology-aware KG-RAG framework compared to existing methods. This work enables reliable knowledge representation for intertwined industrial conditions, contributing to domain-specific RAG development.

published=20251209 Feedback: Relevant | Not Relevant Abstract PDF

Integration of AI-Driven CAD Systems in Designing Water and Power Transportation Infrastructure for Industrial and Remote Landscape Applications

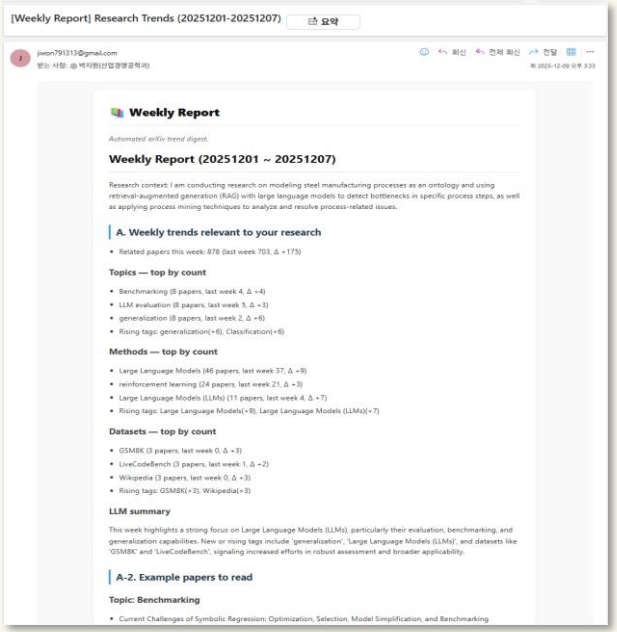
LLM summary

This paper demonstrates AI's application in optimizing industrial design and resource allocation for infrastructure,

일별 연구 동향 브리핑

- arXiv에 올라온 논문 중 관심사와 유사한 Top 4 논문 요약 제공
- LLM 기반 관련 연구의 핵심 Topic, Method, 추천 이유 제공
- 제공된 논문에 좋아요 표시를 남겨 해당 논문을 플랫폼에 저장하고 이를 활용하여 추후 관련 연구를 추천할 예정

#2-01 Mailing - Weekly Report


[Weekly Report] Research Trends (20251201-20251207)

jwon791313@gmail.com

Weekly Report

Automated arXiv trend digest.

Weekly Report (20251201 ~ 20251207)

Research context: I am conducting research on modeling steel manufacturing processes as an ontology and using retrieval-augmented generation (RAG) with large language models to detect bottlenecks in specific process steps, as well as applying process mining techniques to analyze and resolve process-related issues.

A. Weekly trends relevant to your research

- Related papers this week: 878 (last week 703, Δ +175)

Topics — top by count

- Benchmarking (8 papers, last week 4, Δ +4)
- LLM evaluation (8 papers, last week 5, Δ +3)
- generalization (8 papers, last week 2, Δ +6)
- Rising tags: generalization(+6), Classification(+6)

Methods — top by count

- Large Language Models (46 papers, last week 37, Δ +9)
- reinforcement learning (24 papers, last week 21, Δ +3)
- Large Language Models (LLMs) (11 papers, last week 4, Δ +7)
- Rising tags: Large Language Models(+9), Large Language Models (LLMs)(+7)

Datasets — top by count

- GSM8K (3 papers, last week 0, Δ +3)
- LiveCodeBench (3 papers, last week 1, Δ +2)
- Wikipedia (3 papers, last week 0, Δ +3)
- Rising tags: GSM8K(+3), Wikipedia(+3)

LLM summary

This week highlights a strong focus on Large Language Models (LLMs), particularly their evaluation, benchmarking, and generalization capabilities. New or rising tags include 'generalization', 'Large Language Models (LLMs)', and datasets like 'GSM8K' and 'LiveCodeBench', signaling increased efforts in robust assessment and broader applicability.

A-2. Example papers to read

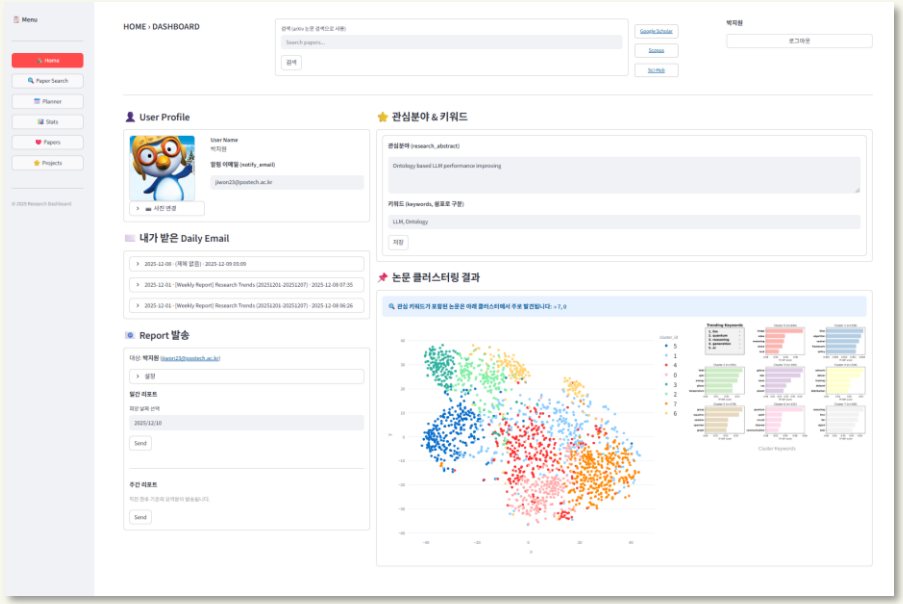
Topic: Benchmarking

- Current Challenges of Symbolic Regression: Optimization, Selection, Model Simplification, and Benchmarking

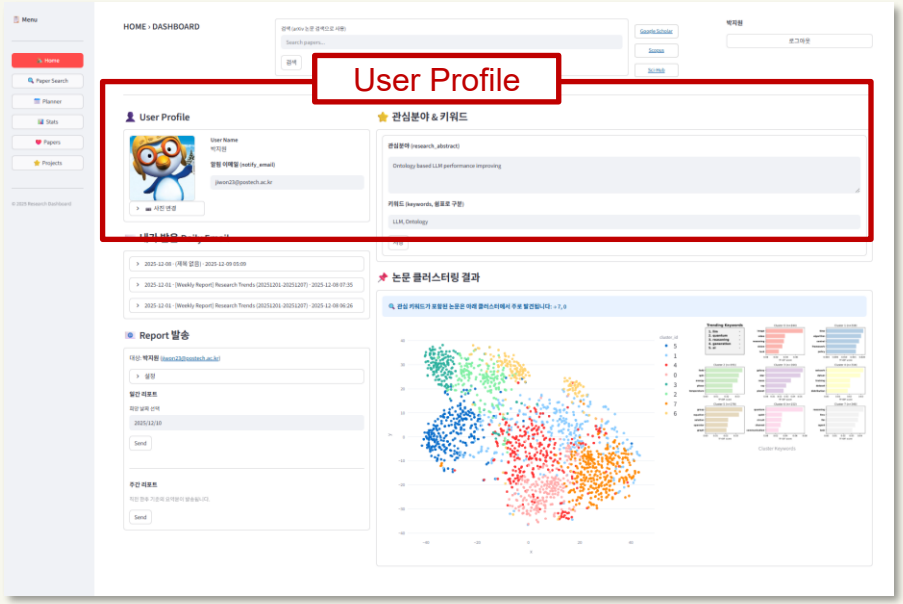
주간 연구 동향 브리핑

- 한 주 동안의 연구 트렌드를 Topic, Method , Dataset 중심으로 요약해 연구 흐름을 파악하도록 돕는 Report
- Part A : 본인 연구 관심사 기반 개인화된 주간 요약 제공
- Part B : 주간 출판된 연구 전체를 기반으로 거시적인 연구 trend 제공

#2-02 개인 대시보드 - Home



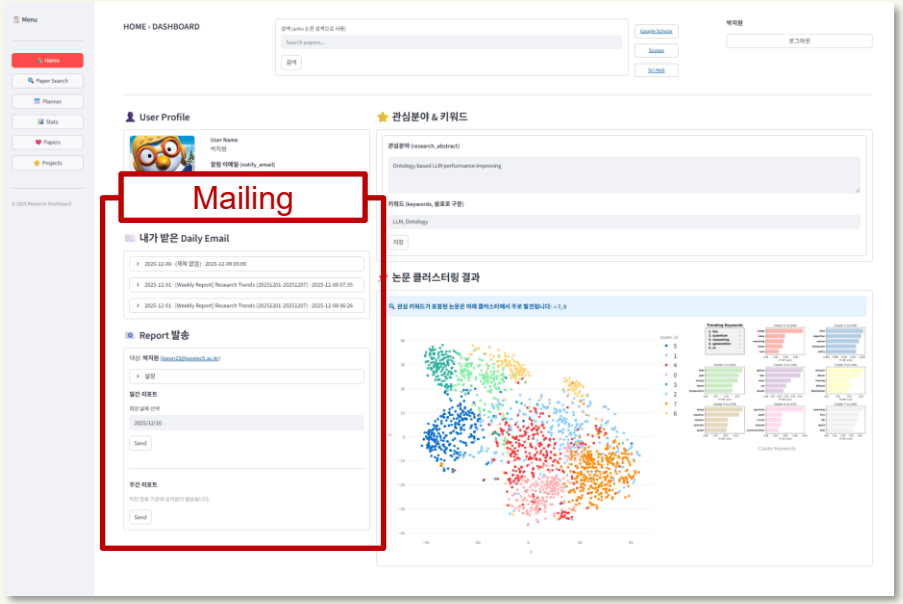
#2-02 개인 대시보드 - Home



User Profile

- 사용자가 본인의 연구분야와 핵심 키워드를 입력하고 관리하는 공간
- 해당 입력은 User DB의 Research_contexts 와 Users와 table에 저장됨

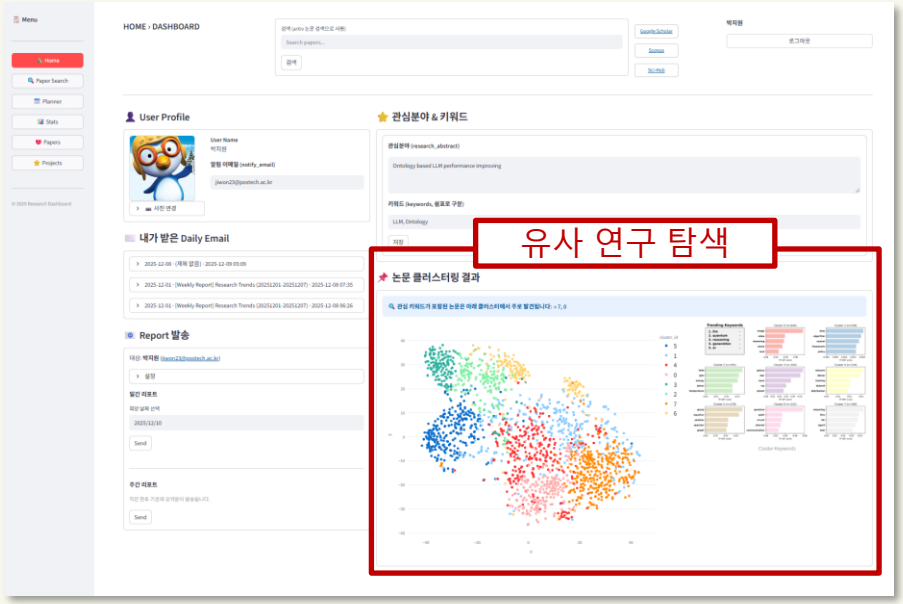
#2-02 개인 대시보드 - Home



Mailing

- 자동 생성된 daily/weekly 리포트 수신 기록을 바로 확인 가능한 공간
- Report 발송 기능을 통해 사용자가 원하는 일자의 arXiv 논문 기반의 요약본을 즉시 받아볼 수도 있음

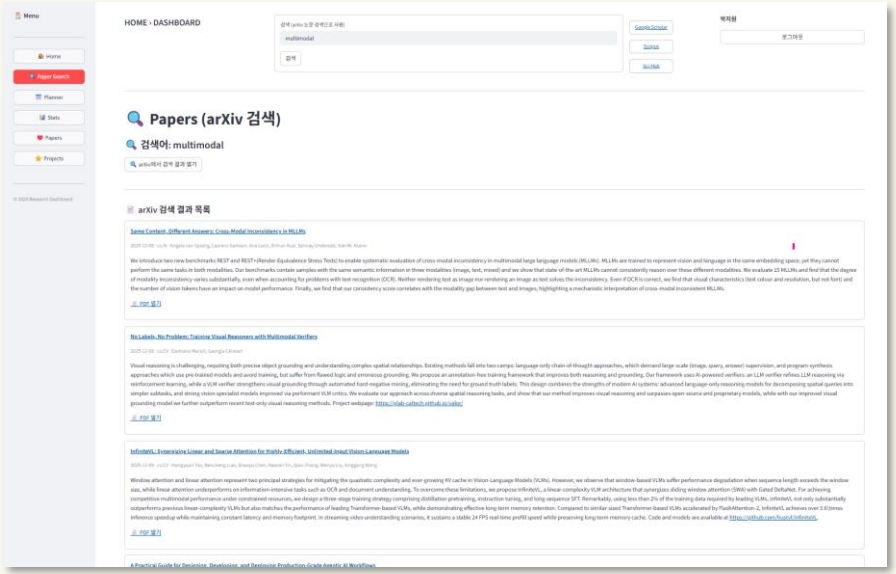
#2-02 개인 대시보드 - Home



유사 연구 탐색

- 관심 연구가 전체 연구 지도에서 어디에 위치하는지 시각적으로 확인
- 최근 자주 등장하는 키워드와 클러스터 별 대표 키워드 제공
- 논문의 title, abstract를 기반으로 k-means Clustering 수행

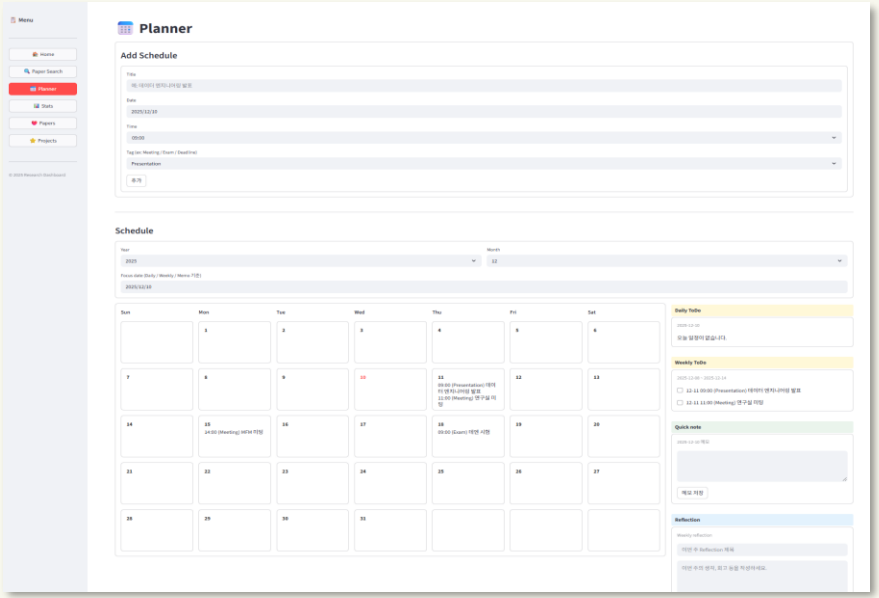
#2-02 개인 대시보드 - Paper Search



Paper Search

- arXiv 내 논문 검색 기능
- 원클릭 외부 연동을 통해, arXiv에 관련 논문이 없을 경우 Google Scholar, Scopus, Sci Hub 등 기타 사이트로 넘어갈 수 있도록 연결

#2-02 개인 대시보드 - Planner



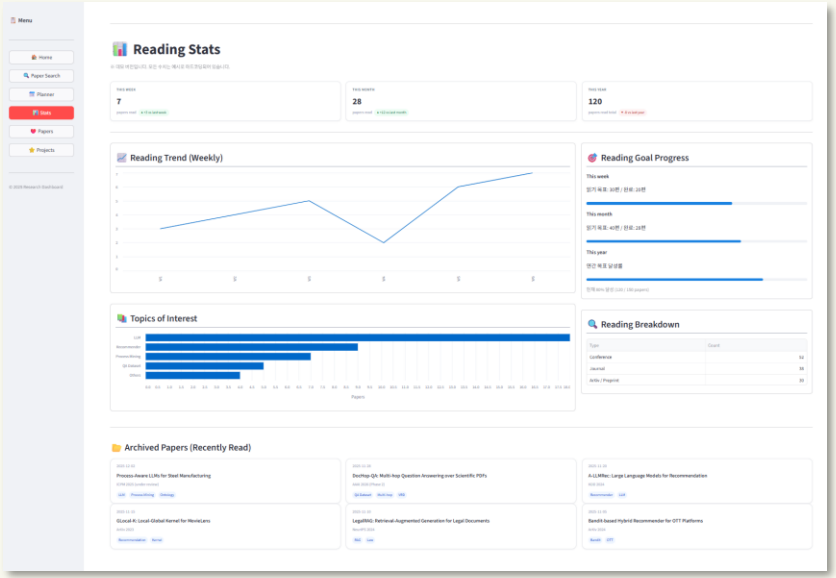
Planner

월간 캘린더 (Monthly Calendar)
주요 미팅, 발표, 마감일 등 연구 일정을 시각적으로 관리합니다.

Todo & Reflection
매일, 매주 해야 할 일을 추적하고, 연구 과정을 회고하며 다음 단계를 계획합니다.

Quick Memo
순간적으로 떠오르는 아이디어나 메모를 빠르게 기록합니다.

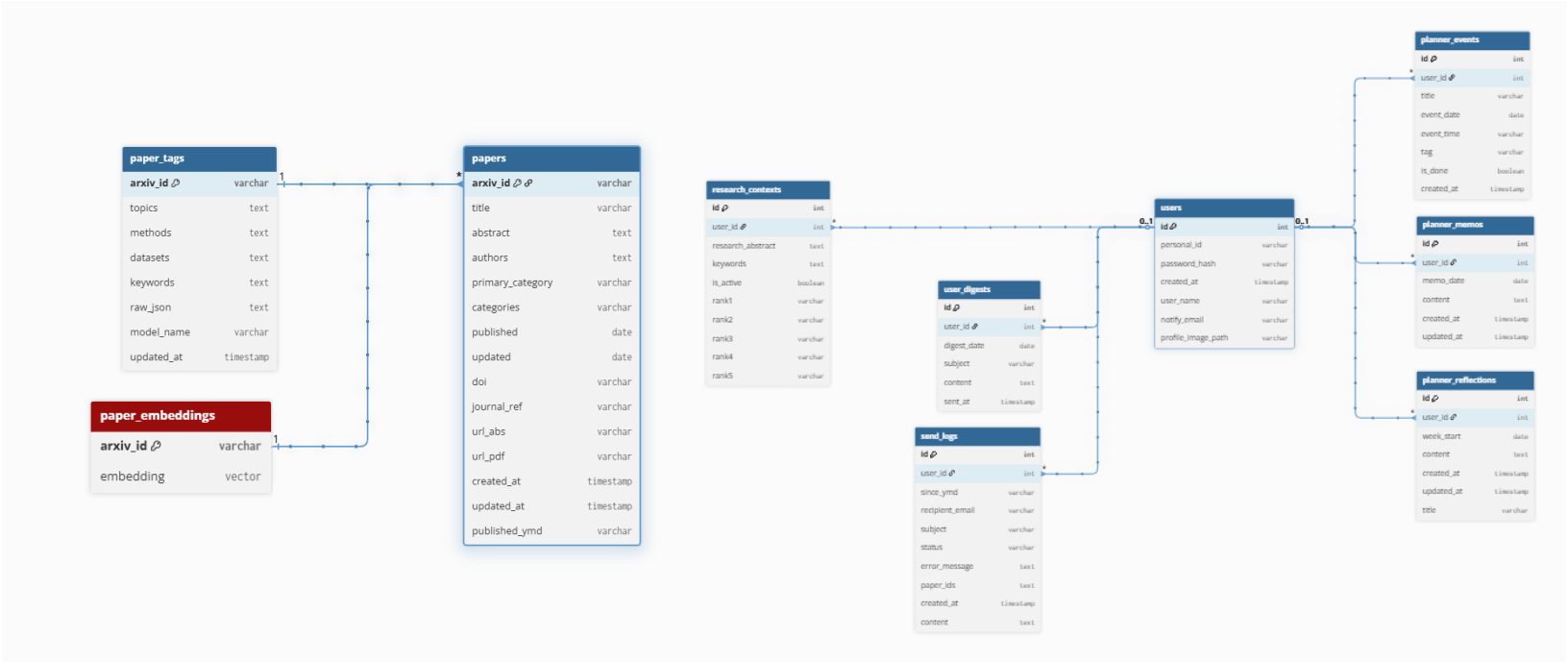
#2-02 개인 대시보드 - Stats



Reading Stats

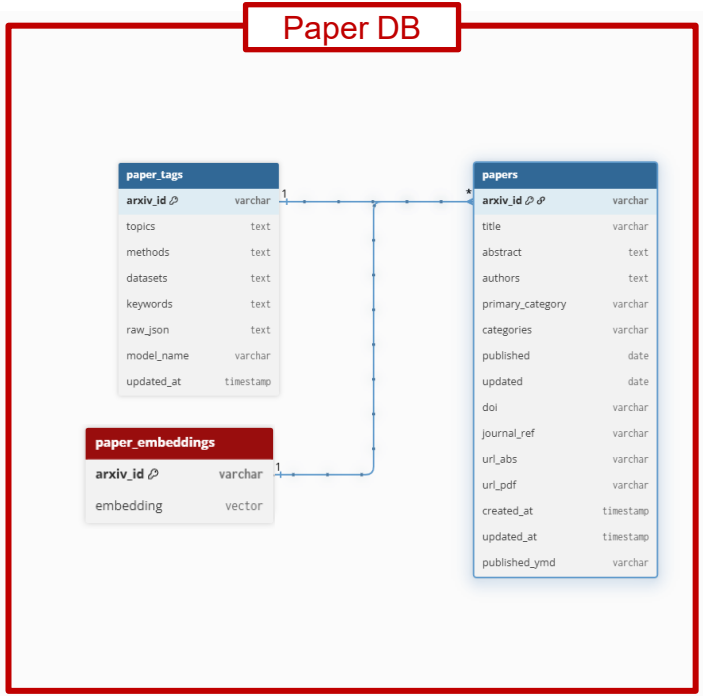
- 주간/월간 읽은 논문 추이를 정량적으로 분석
- 최근에 읽은 논문들을 자동으로 아카이빙하는 기능 제공
- 해당 페이지에서 수집한 데이터를 기반으로 추후 논문 추천이나 관심연구 분야 Update에 활용

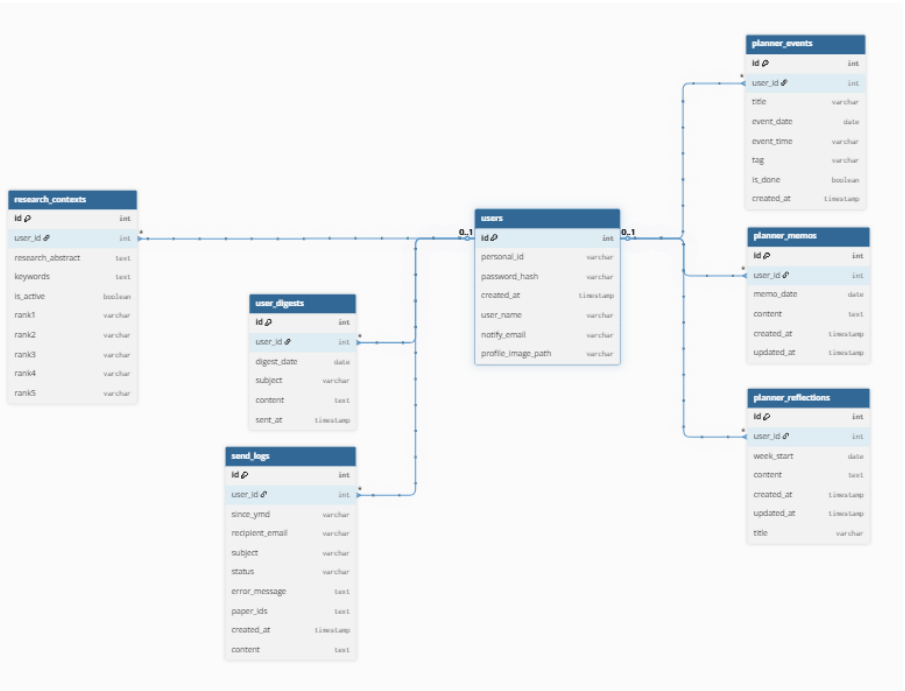
#3-01 DB 구성


paper_tags
arxiv_id varchar
topics text
methods text
datasets text
keywords text
raw_json text
model_name varchar
updated_at timestamp
paper_embeddings
arxiv_id varchar
embedding vector
papers
arxiv_id varchar
title varchar
abstract text
authors text
primary_category varchar
categories varchar
published date
updated date
doi varchar
journal_ref varchar
url_abs varchar
url_pdf varchar
created_at timestamp
updated_at timestamp
published_ymd varchar
research_contexts
id int
user_id int
research_abstract text
keywords text
is_active boolean
rank1 varchar
rank2 varchar
rank3 varchar
rank4 varchar
rank5 varchar
user_digests
id int
user_id int
digest_date date
subject varchar
content text
sent_at timestamp
send_logs
id int
user_id int
since_ymd varchar
recipient_email varchar
subject varchar
status varchar
error_message text
paper_ids text
created_at timestamp
content text
users
id int
personal_id varchar
password_hash varchar
created_at timestamp
user_name varchar
notify_email varchar
profile_image_path varchar
planner_events
id int
user_id int
title varchar
event_date date
event_time varchar
tag varchar
is_done boolean
created_at timestamp
planner_memos
id int
user_id int
memo_date date
content text
created_at timestamp
updated_at timestamp
planner_reflections
id int
user_id int
week_start date
content text
created_at timestamp
updated_at timestamp
title varchar

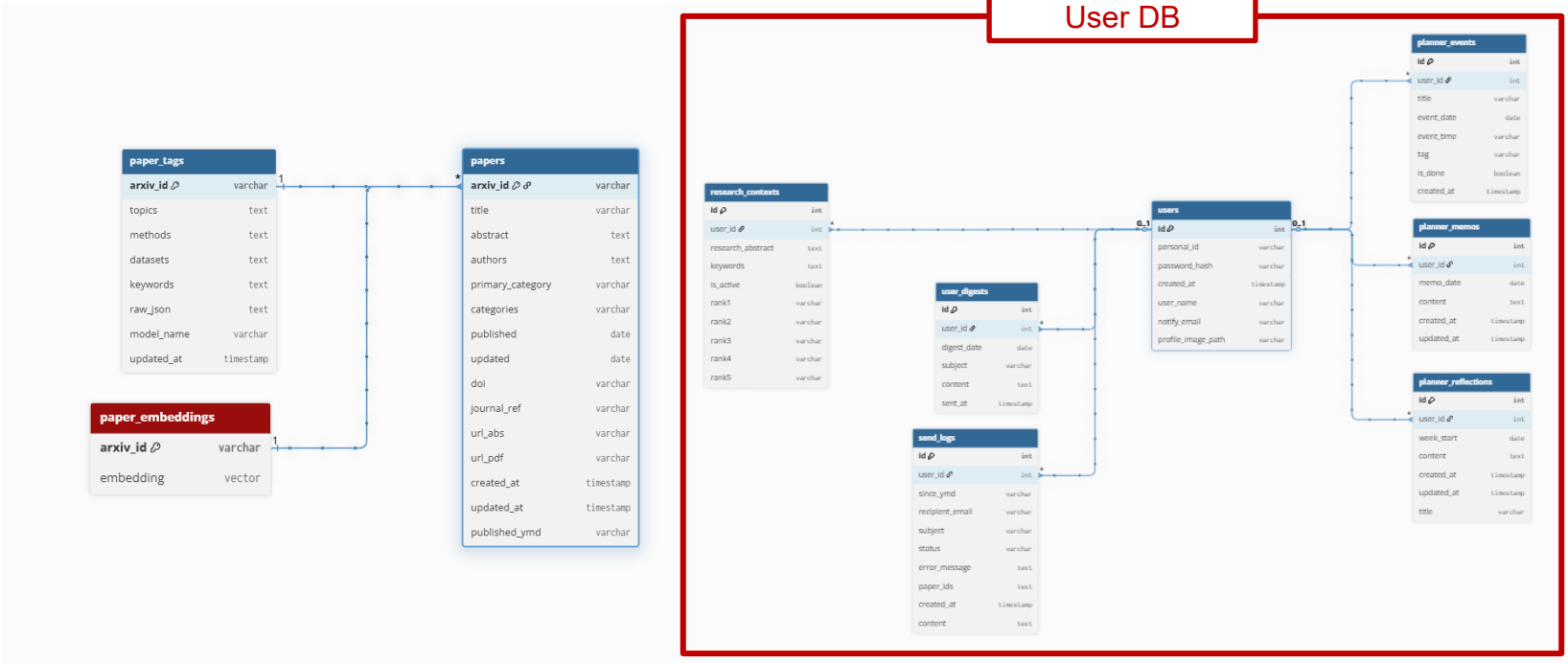
#3-01 DB 구성

Paper DB





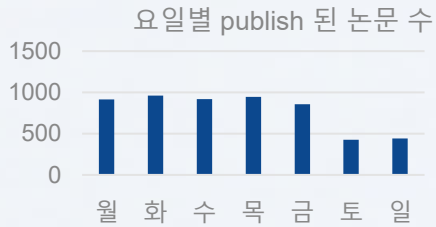
#3-01 DB 구성


User DB
paper_tags
arxiv_id varchar
topics text
methods text
datasets text
keywords text
raw_json text
model_name varchar
updated_at timestamp
paper_embeddings
arxiv_id varchar
embedding vector
papers
arxiv_id varchar
title varchar
abstract text
authors text
primary_category varchar
categories varchar
published date
updated date
doi varchar
journal_ref varchar
url_abs varchar
url_pdf varchar
created_at timestamp
updated_at timestamp
published_ymd varchar
research_contexts
user_digests
send_logs
users
planner_events
planner_memos
planner_reflections

#3-02 Data & Technology

DATA

- **Period :** '25. 09.29 ~ 12.09
- **# of papers :** 48,721 개



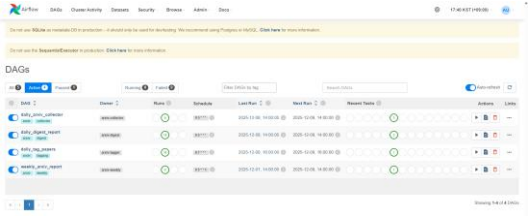
- **Top Category**

  : cs.CV, cs.LG, cs.CL, quant-ph, cs.AI, cs.RO, cs.CR etc

Technology Stack

- **Automation :** Airflow
- **Backend & APIs :** Python, arXiv API, Google gemini API
- **Frontend :** Streamlit
- **Embedding Model :** 'allenai/specter2_base
- **Database :** PostgreSQL, ChromaDB

#3-03 Mailing Airflow



Daily_arXiv_Collector

- Collect_Papers
- Embed_Papers
- Write_Success_Flag

Daily_Digest_Report

- Wait_for_collector_success_flag
- Send_Daily_abstract_digest

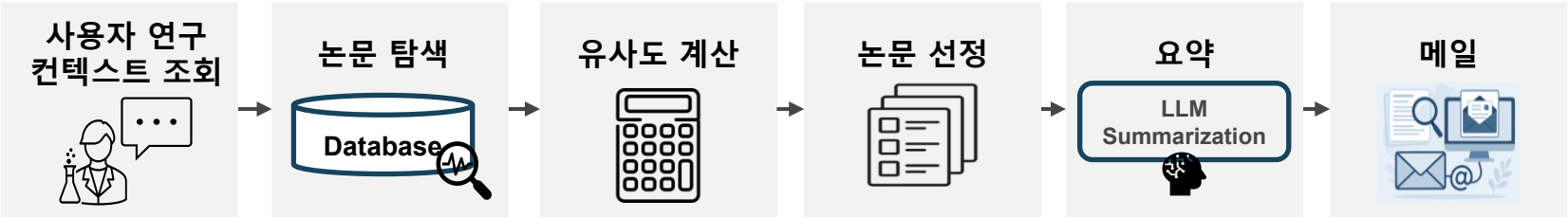
Daily_Tag_Papers

- LLM 기반으로 당일 수집한 논문에 대한 KEYWORD 등을 Tagging하는 작업
- 해당 Tag는 Keyword Trend 파악 등에 활용

Weekly_arXiv_Report

- arXiv가 update 되지 않는 월요일에는 주간 Keyword 트렌드 등을 계산하여 report 생성

#3-04 Daily Report Generation




```
system_prompt = (
    "You are a helpful research assistant. Analyze the paper based on the user's research focus. "
    "Output valid JSON only, with the following keys:\n"
    "- relevance_comment: One concise sentence (max 30 words) explaining why this paper is worth reading. "
    "If similarity_score < threshold, start with 'Loosely related:'.\n"
    "- compressed_summary: A 3-4 sentence summary of the abstract, capturing the core problem, method, and result.\n"
    "- topics: A list of 2-3 key research topics (e.g., 'RAG', 'LLM').\n"
    "- methods: A list of 2-3 key methods or models used (e.g., 'HyDE', 'Transformer')."
)
```

#3-04 Weekly Report Generation




```
system = (
    "Look at the tag changes and write one bullet per tag with an importance label. "
    "Use exactly one of these labels: '[Not important now]', '[Long-term to watch]', '[Urgent trend]'. "
    "Decide based on relevance/impact to the user's research summary; if irrelevant, choose '[Not important now]'. "
    "Write concise English bullets. Add a blank line between bullets for readability."
)
```

#04 결론

- **본 프로젝트 의의**
  - 연구 정보 과부하(Information Overload) 문제를 구조적으로 해결하는 플랫폼 구축
  - Embedding/ Clustering 기반의 연구 동향 탐지 자동화
  - 연구자 라이프사이클을 통합 관리하는 'Research Operating System' 구현

- **추후 연구**
  - LLM 기반 Personalized Digest 성능 고도화
  - 프로젝트/논문 관리 기능 도입 및 해당 정보 수집을 통한 관련 연구 추천 기능 고도화
  - 단순 현재 연구 trend 예측을 넘어, 향후 연구가 어떤 방향으로 나아가야 할 지를 추천하고 제안하는 기능 연구

THANK YOU

