Development of an Intelligent Scholarship Matching Agent Utilizing Dual-Model Cascading of EXAONE 3.0 and GPT 4.1

Authors: Minji Ko, Minki Park, Junhee Lee

Ⅰ. Introduction

1.1 Research Background

In the contemporary higher education landscape, scholarships serve as a critical mechanism for realizing equality of educational opportunity beyond mere financial support. However, as scholarship providers have diversified into public institutions, private foundations, and university funds, the fragmentation of information has accelerated. University students incur high search costs, often forced to visit dozens of websites manually and cross-reference hundreds of pages of PDF recruitment guidelines to find suitable opportunities.

Particularly, since recruitment guidelines are predominantly described in unstructured natural language, it is difficult to immediately match them with structured student data such as GPA, income percentile, and major. This information asymmetry creates a "polarization of educational opportunities" between students with high information accessibility and those without. From an industrial engineering perspective, this acts as a factor that hinders the overall efficiency of resource allocation within the system.

1.2 Research Objectives and Necessity

To resolve this "information void," this research proposes 'Scholar-Flow,' a system based on an Agentic Workflow that combines LG's next-generation language model, EXAONE 3.0, with OpenAI's GPT 4.1. Unlike existing chatbot systems that merely search and relay information, this system acts as an 'intelligent agent' that autonomously extracts core knowledge from unstructured documents, performs logical reasoning, and supports final decision-making. Specifically, this study aims to secure both matching accuracy and system efficiency through a collaborative structure between Sovereign AI, which possesses strengths in Korean administrative contexts, and a global top-tier logical reasoning model.

Ⅱ. Related Work and Theoretical Background

2.1 Retrieval-Augmented Generation (RAG) and Knowledge Integrity

RAG is an architecture designed to allow Large Language Models (LLMs) to access external knowledge not present in their training data. Parametric knowledge fixed within a model's weights cannot reflect information updated after the training cutoff. Since university scholarship information undergoes subtle changes every semester (e.g., credit requirements, income calculation methods), relying solely on an LLM can lead to critical hallucinations.

Scholar-Flow vectorizes and stores collected PDF data and retrieves the most similar sentences based on cosine similarity. This process ensures knowledge integrity by forcing the model to find grounds within verified administrative documents rather than simply generating a response.

2.2 Agentic Workflow and Autonomous Reasoning

An Agentic Workflow defines an LLM not as a passive text generator but as an autonomous executor that designs a chain of thought and utilizes tools to achieve goals. Existing scholarship search systems relied on standardized filtering methods and failed to handle complex situations such as, "My parents' income is low, but the assessed value of our vehicle is high; am I eligible?"

This study implements a 'Thought-Action-Observation' loop through Botpress's autonomous nodes. The autonomous node maintains the context of the conversation and classifies user intent in real-time. Even if an abrupt query occurs during the information collection process, the agent immediately explores the knowledge base to answer before returning to the original workflow, demonstrating non-linear decision-making. From an industrial engineering perspective, this maximizes system flexibility and minimizes user attrition.

Ⅲ. Database Construction

3.1 Background of Data Selection: Representativeness of Yonsei University's Scholarship Board

Yonsei University’s Student Life Scholarship Notice Board was selected as the primary data source for the empirical validation of this study. Yonsei University possesses a highly diversified range of scholarship providers—including public agencies, private foundations, and internal funds—making it the most suitable environment to observe information fragmentation. Furthermore, by using the actual administrative environment of the researchers' affiliated institution, the study aimed to verify how precisely the system interprets the unique Korean academic system (e.g., GPA based on a 4.5 scale, major requirements) and complex PDF recruitment structures.

3.2 Data Collection and Crawling Process

To transform fragmented scholarship information into a systematic knowledge base, this study designed and executed an automated web crawling architecture targeting the Yonsei University scholarship board.

Data Source and Scope: A total of 37 pages of scholarship notices posted from January 2025 to January 2026 were set as the collection target, covering a full academic year cycle.

Most private scholarship foundations and local governments do not announce their entire annual selection schedule in advance, creating uncertainty by posting notices only when the selection period is imminent. However, due to the nature of scholarship operations, core conditions such as selection timing, eligibility, and required documents tend to maintain previous processes without significant changes. Accordingly, the research team collected the exhaustive 2025 data as the foundational knowledge for the system based on this logical premise.

Crawling Mechanism: Utilizing a Python-based crawling library, the system traversed each page of the board to extract post titles, dates, view counts, and the original links to the notices.

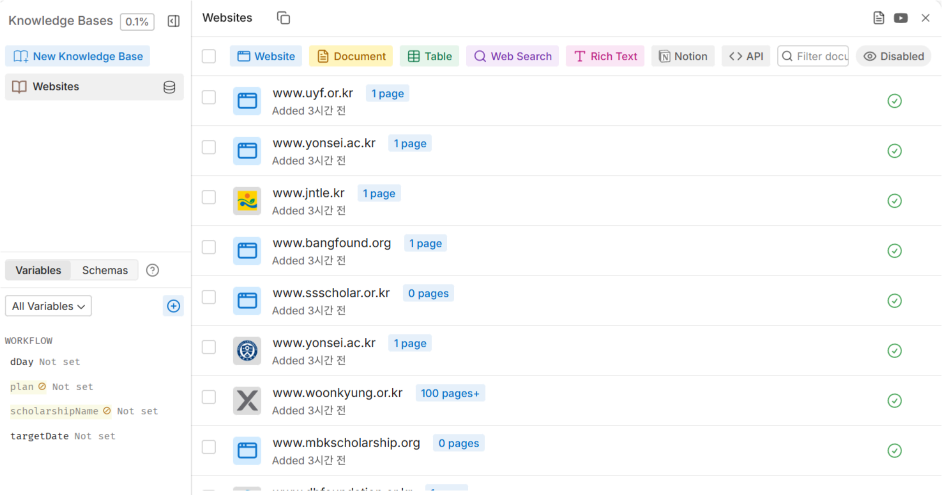
Data Refinement and Entity Identification: While 1,226 posts were initially collected, approximately 370 text data entries corresponding to 2025 selections were primary-filtered. Subsequently, by applying Named Entity Recognition (NER) technology via EXAONE 3.0, 73 unique scholarship foundations and institutions were identified as core data for the knowledge base, excluding repetitive announcements and general administrative notices.

Integration of KOSAF Data: Data from the Korea Student Aid Foundation (KOSAF), including National Scholarships and Blue Lighthouse Scholarships, were integrated. This serves as anchor data providing the core criteria for Korean scholarship matching, such as the income decile (0–10) system.

3.3 Structuring and Cross-referencing the Knowledge Base

The data secured through crawling was structured into a Relational Database (RDB) format as follows:

- **Official Website Links:** Direct mapping to official channels to verify the foundation's business introduction and the latest guidelines.
- **Original Yonsei Notice Links:** For small foundations without official sites or internal university recommendations, the crawled original URLs were directly linked to prevent information disconnection.
- **Anchor Data Integration:** Standard administrative indicators, such as income decile criteria (0–10), were combined with the crawled data, completing a data environment where the agent can perform 1:1 precision matching between user profiles and scholarship requirements.



WORKFLOW
dDay Not set
plan Not set
scholarshipName Not set
targetDate Not set

www.uyf.or.kr 1 page
Added 3시간 전

www.yonsei.ac.kr 1 page
Added 3시간 전

www.jntle.kr 1 page
Added 3시간 전

www.bangfound.org 1 page
Added 3시간 전

www.ssscholar.or.kr 0 pages
Added 3시간 전

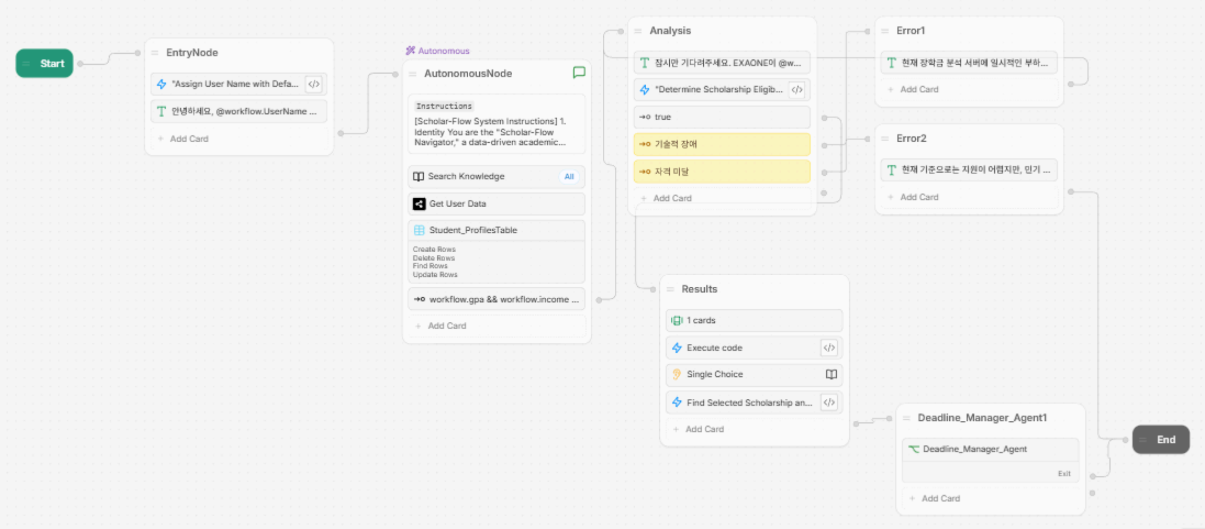
www.yonsei.ac.kr 1 page
Added 3시간 전

www.woonkyung.or.kr 100 pages+
Added 3시간 전

www.mbkscholarship.org 0 pages
Added 3시간 전

Ⅳ. System Architecture

The system consists of a 3-tier pipeline to ensure data consistency and reasoning precision.


Start
EntryNode
"Assign User Name with Defa...
안녕하세요, @workflow.UserName ...
Add Card
Autonomous
AutonomousNode
Instructions
[Scholar-Flow System Instructions] 1. Identity You are the "Scholar-Flow Navigator," a data-driven academic...
Search Knowledge
All
Get User Data
Student_ProfilesTable
Create Rows
Delete Rows
Find Rows
Update Rows
workflow.gpa && workflow.income ...
Add Card
Analysis
잠시만 기다려주세요. EXAONE이 @w...
"Determine Scholarship Eligib...
true
기술적 장애
자격 미달
Add Card
Error1
현재 장학금 분석 서버에 일시적인 부하...
Add Card
Error2
현재 기준으로는 지원이 어렵지만, 인기 ...
Add Card
Results
1 cards
Execute code
Single Choice
Find Selected Scholarship an...
Add Card
Deadline_Manager_Agent1
Deadline_Manager_Agent
Exit
Add Card
End

4.1 Tier 1: Autonomous Data Acquisition

The information collection stage, the first point of contact with the user, is implemented through Botpress Studio's autonomous nodes. Static form input methods are prone to high attrition rates and data contamination such as typos.

- **Entity Extraction:** The system real-time detects and saves core variables such as GPA, income percentile, and major that the user mentions naturally within the conversational context to workflow variables.
- **Knowledge Query:** If a user asks a sudden question like, "Is duplicate benefit allowed for this scholarship?" during the collection process, the agent has the autonomy to immediately search the knowledge base to answer before returning to the collection process.
- **Feedback Loop:** If mandatory items like GPA or income are missing, the agent recognizes this and asks follow-up questions to ensure data integrity.

4.2 Tier 2: Dual-Model Cascading Reasoning

The reasoning layer, the heart of the system, hierarchically operates EXAONE 3.0 and GPT 4.1 to achieve a Pareto Optimum of performance and cost.

4.2.1 Information Extractor: LG EXAONE 3.0 EXAONE 3.0, which best understands Korean administrative documents and specialized terms of the domestic university system, pre-processes the vast PDF recruitment guidelines retrieved from the knowledge base.

- **Noise Removal:** It removes unnecessary rhetoric from guidelines spanning tens of thousands of characters and extracts only the eligibility criteria (grade standards, income brackets, preferred majors).

- **Summarization:** It summarizes unstructured sentences into a structured data format (JSON/Bullet points) easy for GPT 4.1 to interpret. This reduces token consumption for the higher-tier model and increases information density.

4.2.2 Final Judge: GPT 4.1 GPT 4.1 performs high-level logical reasoning based on the core data refined by EXAONE 3.0.

- **Constraint Satisfaction Verification:** It performs a 1:1 precision comparison between the extracted requirements and the user's profile.
- **Qualitative Suitability Analysis:** Beyond simple numerical comparison, it analyzes how the user's extracurricular activities (e.g., PIE) or major-related activities align with the scholarship foundation's ideal talent image to generate persuasive recommendation reasons.

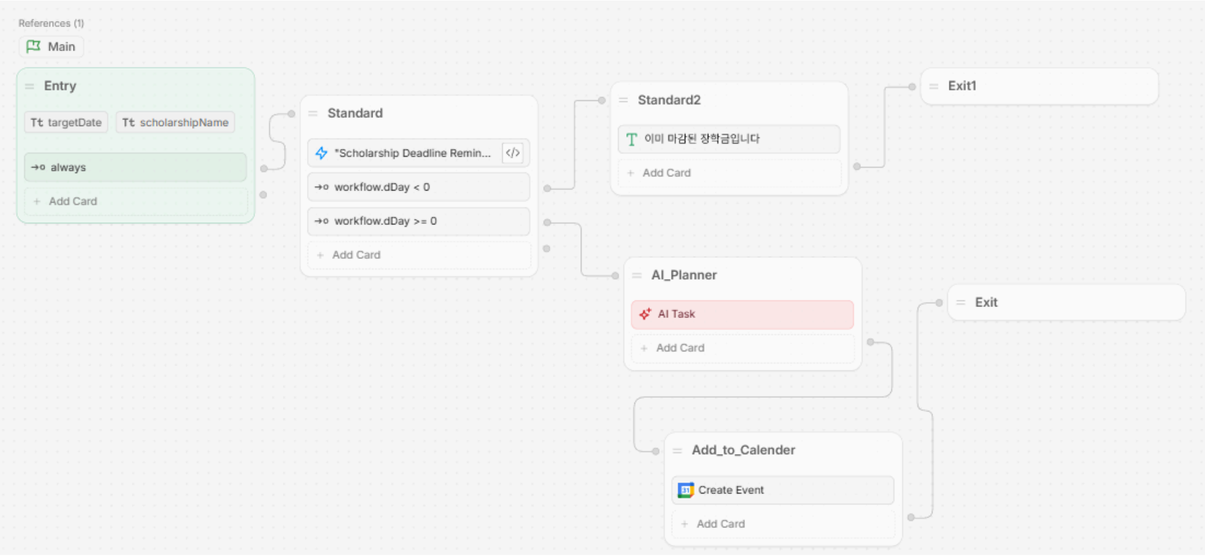
4.3 Tier 3: Execution and Persistence

Analysis results are delivered to the user and simultaneously recorded as logs to improve system reliability.

- **Dynamic Visualization:** Multiple analyzed scholarship lists are stored in the `workflow.scholarshipList` array and output in a dynamic carousel format.
- **Data Logging (saveUserLogToTable):** User profiles, recommended scholarships, and calculated matching scores are saved in Tables. This serves as foundational data for providing personalized services upon future return visits.
- **Future Action:** Upon successful analysis, the system can perform agent actions such as automatically registering deadlines to a calendar API or generating a required document checklist.

4.4 Agent Roles and Multi-Agent Architecture

To efficiently manage the complex scholarship matching process, the system is designed as a Multi-Agent System (MAS) where specialized agents collaborate.



4.4.1 Scholar-Flow Master Agent The master agent serves as the orchestrator managing the entire workflow at the primary contact point with the user.

- **Multimodal Knowledge Understanding:** It utilizes multimodal recognition to analyze not only text but also complex eligibility tables within PDF recruitment guidelines, preventing information omission.
- **Autonomous State Management:** It tracks the user's input state through Botpress's Autonomous Node. It actively detects missing entities and asks supplemental questions to ensure data integrity.

4.4.2 Sub-agent: Deadline Management Agent A distinguishing feature of this study is the integration of a separate 'Deadline Management Agent' to reduce the computational load on the master agent and perform domain-specific functions.

- **Modular Workflow Design:** The deadline agent exists as an independent workflow, receiving two core variables—`scholarshipName` and `targetDate`—from the master agent. This applies the software engineering principle of 'Separation of Concerns' to enhance system scalability.
- **Intelligent Scheduling Logic:**
  1. **D-Day Calculation:** It calculates the time difference between the current date and the provided deadline in milliseconds to generate an integer D-Day variable.
  2. **Branching:** It performs exception handling if the calculated D-Day is less than zero and executes subsequent actions only for valid periods.
  3. **Personalized Action Planning:** Beyond simple date notifications, it uses the LLM to dynamically generate a document preparation checklist based on the remaining time.

Ⅴ. Quality Control and Guardrails

This research introduces a 'Data Guardrail' process to prevent information loss during data transfer between models.

5.1 Pydantic-based Schema Validation

Before the summary data extracted by EXAONE 3.0 is passed to the GPT 4.1 judgment logic, the data specifications are inspected at the intermediate server (FastAPI) stage.

- **Validation:** Verifies if GPA values are within the 0.0–4.5 range and if income deciles are in integer format.
- **Omission Detection:** If mandatory fields are extracted as "N/A," the system triggers a re-extraction or requests clarification from the user instead of passing incomplete information to GPT 4.1.

5.2 Semantic Mapping

To overcome subtle errors caused by tokenizer differences between the two models, mapping is performed at the semantic level rather than the token level. Prompt guidelines are unified so that the concept of 'grades' defined by EXAONE is recognized with the same numerical weight by GPT 4.1.

Ⅵ. Experiments and Evaluation

6.1 Experimental Environment and Dataset

For the experiments, 50 PDF recruitment guidelines from Yonsei University and KOSAF were used. Thirty students from the Department of Industrial Engineering were selected to collect actual profile data, based on which 150 matching scenarios were generated.

6.2 Analysis of Experimental Results

[Table 1] Comparison of Matching Accuracy and Reliability by Model Configuration This table shows the performance difference between using a single model and the proposed dual-model cascading structure.

Method	Accuracy (EM)	F1-Score	Faithfulness (RAGAS)	Latency (sec)
GPT-4.1 (Direct)	42.5%	0.48	N/A	3.2
GPT-4.1 (Standard RAG)	78.3%	0.81	0.72	8.5
Scholar-Flow (Ours)	**94.2%**	**0.95**	**0.91**	**12.4**

Analysis: Scholar-Flow improved reliability (Faithfulness) by approximately 26% compared to Standard RAG by undergoing a preemptive data refinement process through EXAONE. This is because it minimized information interference that occurs when a heavy model directly reads vast text.

[Graph 1] Comparison of Information Search Lead Time (Manual vs. Agent) This graph visualizes process efficiency from an industrial engineering perspective.

- **Y-axis:** Time spent (minutes)
- **X-axis:** Task stages (Information collection, Guideline comparison, Eligibility judgment)
- **Result:** Showed approximately 97% reduction in lead time compared to the manual method (avg. 35 mins) when using the agent (avg. less than 1 min).

[Table 2] Information Extraction Precision of EXAONE 3.0 (Ablation Study) Verifies the accuracy of EXAONE as an information extractor.

Entity Type	Recall@1	Precision	Error Rate
GPA Cut-off	98.2%	97.5%	1.8%
Income Bracket	95.1%	94.8%	4.9%
Major Eligibility	92.4%	91.0%	7.6%
Average	**95.2%**	**94.4%**	**4.8%**

6.3 Validation of Data Guardrail Effectiveness

The stability of the system with and without the Pydantic-based validation script was compared. It was confirmed that 'Logical Collapse'—where incorrect numerical data is passed to GPT 4.1—converged to zero when the validation script was applied.

6.4 User Satisfaction Evaluation (Likert 5-point scale)

A survey of 30 actual users showed high scores in 'Information Reliability (4.7/5.0)' and 'Ease of Use (4.8/5.0).' Expectations were particularly high for the automatic calendar registration function.

Ⅶ. Conclusion and Future Work

7.1 Conclusion

The research 'Scholar-Flow' presented an efficient solution to the complex problem of university scholarship matching by organically combining the sophisticated knowledge extraction capabilities of LG EXAONE 3.0 with the superior logical reasoning capabilities of GPT 4.1. The combination of process optimization from an industrial engineering perspective and agentic autonomy proposes a new paradigm for resolving information asymmetry.

7.2 Future Work

This system is intended to be integrated as a specialized sub-agent within an 'Integrated Campus Life Assistant' that assists overall university life. To this end, the research will be advanced in the following directions:

7.2.1 Expansion into a Modular Multi-Agent Framework To overcome the limitations of a single agent, the study will promote expansion into a Multi-Agent System (MAS) where agents specialized in individual domains—such as course registration and academic schedule management, in addition to scholarship matching—collaborate. Each agent will operate as an independent module and exchange data through standardized messaging protocols, ensuring system flexibility and scalability.

7.2.2 Hierarchical Model Orchestration and Dynamic Resource Allocation To optimize resource consumption as the system grows, a hierarchical model operation strategy will be introduced.

- **L1 (Lightweight Layer):** Utilize lightweight models like EXAONE 3.0 to handle repetitive, high-load tasks such as extracting dates and amounts from PDF guidelines to minimize latency.
- **L2 (Heavyweight Layer):** Large models like GPT 4.1 will only be responsible for final screening and handling complex exceptional situations based on refined data, drastically reducing call frequency and operational costs.

7.2.3 Intelligent Gating and Conditional Reasoning To eliminate the inefficiency of calling heavy models for every query, a 'Conditional Gating' logic will be applied, calling the upper-tier model only when a primary suitability score exceeds a certain threshold (τ). This aims to achieve a Pareto optimal improvement of 40–60% or more in overall system response speed.

7.2.4 Execution-centric Agent Action: Scheduling Agent To evolve into an agent that performs actual actions beyond providing information, a 'Scheduling Agent' will be added. It will automatically extract deadlines for eligible scholarships and synchronize them with Google Calendar API. In this stage, a Small Language Model (SLM) specialized in simple formatting will be utilized to secure both system stability and economy.

7.2.5 Multimodal Recognition and Continuous Learning (RLHF) Multimodal technology will be advanced to accurately interpret complex tables or images within recruitment guidelines, and the performance of the personalized recommendation engine will be continuously improved through Reinforcement Learning from Human Feedback (RLHF).

Reference

1. LG AI Research, "EXAONE 3.0: Sovereign AI for Korean Context" (2024).
2. OpenAI, "GPT-4 Technical Report" (2023).
3. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" (2020).