# hoBIT: A Profile-Aware Retrieval-Augmented Chatbot for University Academic Advising

**Yoonseo Kim** **Seongmin Lee** **Joongheon Kim** **SeongKu Kang***
Korea University
Seoul, Korea
{seo3167,kyne0127,joongheon,seongkukang}@korea.ac.kr

Code Demo Video Homepage

## Abstract

In university academic advising, identical questions can require different answers depending on a student's department, admission cohort, and degree program, causing profile-blind retrievers to surface plausible but inapplicable evidence. We present **proFILL**, a method for transforming **hoBIT**, our college's current rule-based advising chatbot, into a profile-aware retrieval-augmented generation (RAG) system. Rather than requiring a complete user profile upfront, proFILL progressively acquires only the profile attributes needed for each query, guided by both the query intent and the initially retrieved evidence, and uses them to condition retrieval over a profile-aware index. Extensive experiments and a human preference study show that proFILL outperforms diverse RAG baselines, is preferred by target users, and remains effective with open-weight models for cost-effective on-premise deployment.

## 1 Introduction

Retrieval-augmented generation (RAG) is increasingly employed in specialized domains (Gao et al., 2023b), including medical education (Jang et al., 2025) and religious question answering (Gao et al., 2025). University academic advising is a practical and high-impact application. Students frequently ask questions about graduation requirements, required courses, double-major rules, and scholarship eligibility, yet an incorrect answer can directly affect course planning, delay graduation, or increase the administrative workload of advising offices.

The central difficulty is that academic advising questions are often *profile-dependent*. The correct answer is not determined by the query alone, but also by a student's profile, including department, admission cohort, major type, and related attributes. In our setting, students belong to one of three departments: Computer Science (CS), Data Science (DS), and Artificial Intelligence (AI), where academic requirements and rules vary across admission cohorts. Even within the same department, students may follow different rules depending on their major track, such as intensive, double-major, or convergence programs. Thus, the same question may require different reference documents for different students. Conventional RAG struggles in this setting, as semantically similar documents can lead a profile-blind retriever to return a plausible but wrong source (Figure 1). Hand-written FAQ rules are also hard to maintain, as profile-specific cases accumulate with policy revisions over time.

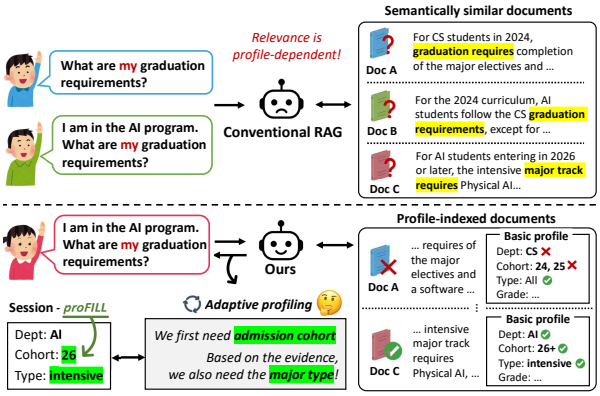


Figure 1: A conceptual comparison of (top) conventional RAG, which struggles with profile-dependent queries over semantically similar documents, and (bottom) the proposed hoBIT, which adaptively acquires the required profile information and retrieves the applicable document from a profile-indexed corpus.

To address this gap, we present proFILL, a method for transforming **hoBIT**, our college's currently deployed rule-based academic advising chatbot, into a profile-aware retrieval-augmented system. Our key idea is to treat profile-dependent evidence validity as part of the index structure, rather than as context to be resolved only at query time. We implement this idea through *offline profile-based indexing*, built on a lightweight academic

* Corresponding author.



*Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 604–613
October 24-29, 2026 ©2026 Association for Computational Linguistics

profile schema covering department, admission cohort, major type, and related attributes. During offline preprocessing, we use this schema to structure institutional materials, including regulations, department webpages, and notices. Each chunk is annotated with the profile values for which it is valid and the profile fields needed to interpret it. This turns the corpus into profile-conditioned evidence, allowing retrieval to distinguish not only what a document is about, but also which students it applies to.

At online query time, we use this indexed structure to acquire the student profile on demand. Instead of requiring login or a complete profile upfront, we maintain a time-limited session profile and collect only the fields needed by the current query. This is achieved via *on-demand adaptive-profiling*: we first infer the query intent, identify the necessary profile fields, and ask only for missing ones. The filled profile is combined with the query to match and filter against the indexed profile-conditioned evidence. We then inspect the retrieved evidence and ask a follow-up question only when fine-grained additional information is needed to verify applicability, such as detailed scholarship eligibility conditions. Once the user provides the additional field, the session profile is updated and retrieval is rerun with the completed profile. This evidence-driven selective process reduces initial user burden by avoiding a full upfront profile form, while improving answer correctness by triggering additional questions only when the retrieved evidence reveals unresolved profile requirements.

**Contributions.** **(1)** We present proFILL, a method for transforming **hoBIT** into a profile-aware RAG system that replaces hand-written FAQ rules with profile-conditioned retrieval. **(2)** In proFILL, we redesign both offline indexing and online query processing to address the profile-dependent nature of academic advising. **(3)** Extensive experiments show that proFILL outperforms diverse RAG baselines, including profile-blind reranking and query-augmentation methods, while remaining effective across various open-weight models.

## 2 The hoBIT System

**Overview.** hoBIT is an academic-advising chatbot deployed at our college of informatics. Its original backend is a rule-driven dialogue system that maps user queries to predefined FAQ responses, limiting its ability to handle the profile-dependent questions common in academic advising. We extend hoBIT with proFILL, a profile-aware retrieval-augmented generation framework featured with two key components: *offline profile-aware indexing* and *online on-demand profiling* (Figure 2).

The resulting system can support a range of advising needs, including curriculum and graduation requirements, notices, and career-related questions. The default profile schema comprises five attributes: `department`, `major type`, `grade`, `admission year` (i.e., admission cohort), and `student status`. These attributes determine which curricula, policies, and other institutional information apply. The schema can be flexibly extended with additional attributes to support new advising needs.

### 2.1 Offline Profile-based Indexing

We build the corpus from five institutional sources: regulation PDFs, orientation PDFs, department webpages, board notices, and administrator-curated FAQs. After cleaning and chunking the collected materials, we use an LLM to annotate each chunk with a structured profile record over the five predefined attributes. Specifically, for each chunk, the LLM identifies which attributes determine the students to whom the information applies and fills in their corresponding values. Attributes that do not affect the chunk's applicability are left `null`. The resulting non-null fields encode the chunk's profile-dependent applicability and unveil which user attributes must be known for reliable retrieval.

In practice, we maintain separate static and dynamic indices based on source update frequency. Relatively stable materials (e.g., regulations) are stored in the static index, whereas frequently updated materials (e.g., notices) are stored in the dynamic index. Results from both indices are combined at query time. Further details on preprocessing and indexing are provided in Appendix A.

### 2.2 Online Query Processing

#### 2.2.1 Intent Routing and Session-Profile Initialization

**Intent Routing.** At query time, an LLM first classifies each incoming query into one of five intents: *greeting*, *ability*, *faq*, *smalltalk*, and *retrieval*. Only *retrieval* queries proceed to subsequent retrieval and answer generation. *greeting*, *ability*, and *faq* queries are handled using predefined templates or the FAQ store without an additional LLM

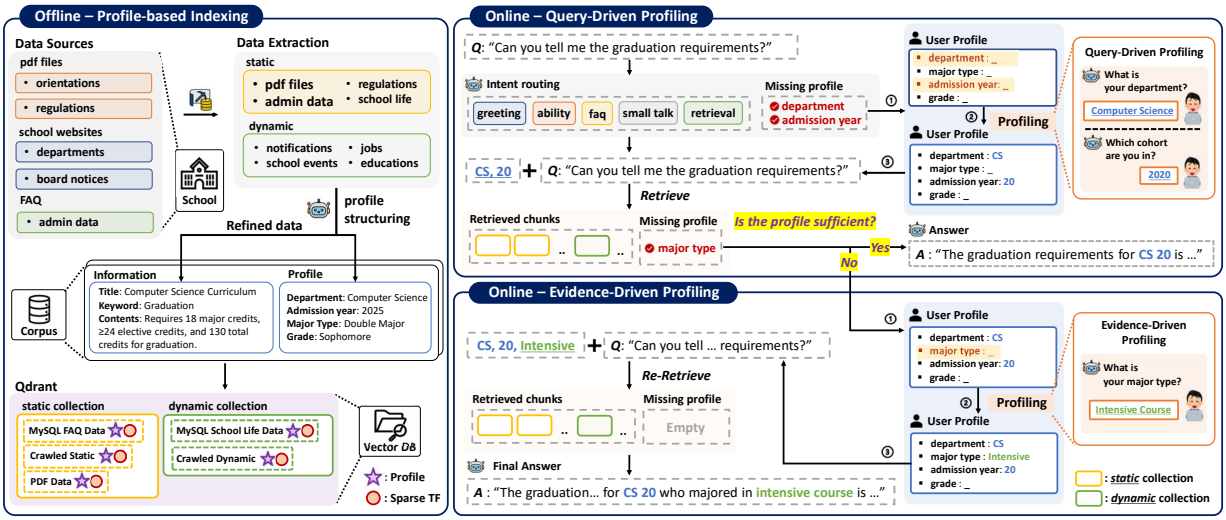


Figure 2: Overview of hoBIT. **Offline (left):** each chunk is annotated with its applicable profile values and required fields before indexing. **Online (right):** query-driven profiling acquires missing attributes for profile-aware retrieval, while evidence-driven profiling requests additional information and re-retrieves when needed.

generation call, while *smalltalk* queries receive a lightweight conversational response.

**Session-Profile Initialization.** For each *retrieval* query, the LLM identifies the profile attributes required to answer the question and extracts any corresponding values available from the query. These values initialize the session profile, while required attributes whose values are unavailable are marked as missing. These missing values are subsequently acquired as needed via adaptive profiling.

Importantly, the initialized profile serves only as a guide rather than a fixed specification of all information required to answer the query. Additional information that cannot be inferred from the query or is not covered by the predefined profile schema may be acquired on demand when its necessity becomes evident from the retrieved evidence.

### 2.2.2 On-demand Adaptive Profiling and Profile-Aware RAG

Rather than requiring login or a complete profile form upfront, proFILL acquires only the information needed for the current query through two complementary stages: *query-driven profiling* and *evidence-driven profiling*.

**Query-Driven Profiling and Retrieval.** Before retrieval, the system asks the user for the attributes marked as missing during session-profile initialization. The acquired values are added to the session profile and incorporated into retrieval through both *soft* query augmentation and *hard* filtering. Specifically, the available profile values are serialized and prepended to the original query for retrieval.[1] The retrieved candidates are then filtered using the profile annotations: Chunks whose specified profile values conflict with the session profile are removed, while chunks with matching or unspecified (null) profile values remain eligible. The top-10 eligible chunks are provided to the generator LLM.

Before generating an answer, an LLM selects the chunks needed to answer the query as the evidence set and checks whether their applicability can be determined from the current session profile. Specifically, it identifies any non-null profile fields of the selected chunks whose corresponding attributes remain unresolved, as these fields indicate the information required to determine whether each chunk applies to the user. The LLM also checks whether additional user information is needed to interpret the evidence reliably. If no further information is required, the system generates an answer from the selected evidence; otherwise, evidence-driven profiling is triggered.[2]

**Evidence-Driven Profiling and Re-Retrieval.** When additional information is required, the system asks the user a targeted follow-up question about the unresolved profile fields or other required information. The acquired information is added to the session profile, and retrieval is repeated in the same manner using the updated profile. An LLM then reselects the relevant evidence and generates the final answer from the refined evidence set.

[1] Further details on retrieval are provided in Appendix B.
[2] In our setup, this occurs for 24.3% of queries on average.

### 2.3 System Implementation

To improve usability and user trust, hoBIT incorporates two interaction design features. First, for attributes covered by the predefined profile schema (e.g., `admission year`), the system presents predefined options instead of requiring free-form input, reducing user effort. Second, each generated answer is accompanied by the selected evidence, allowing users to directly inspect the institutional sources supporting the response.

## 3 Experimental Setup

**Corpus.** We construct the corpus from 515 institutional sources collected from our college: 3 academic-regulation and orientation PDFs, 89 department webpages, 137 board notices, and 286 administrator-maintained FAQ entries.

**Profile-Grounded QA Data.** With guidance from our college's academic affairs office and drawing on query logs from the deployed hoBIT service, we use LLM-assisted generation to construct 1,800 QA instances spanning 60 unique student profiles, 10 advising categories,[3] and three query types: formal, first-person, and affirmative or negative verification queries. Further details on dataset construction and additional experiments for a broader evaluation, including intent routing and open-ended advising, are presented in Appendix C.

**Evaluation Settings.** We evaluate the profile-grounded QA under two settings. **Deployment** reflects the realistic scenario in which the user profile is initially unavailable and must be acquired during interaction. **Oracle** provides the complete profile along with the query, allowing us to assess how effectively proFILL leverages the available profile information. For RAG baselines, the given profile is linearized and appended to the query as additional context. For all methods, we use `text-embedding-3-small` as the dense retriever and `gpt-4o-mini` as the generator. Latency is measured using a 48 GB MIG instance on a single NVIDIA RTX PRO 6000 GPU and an Intel Xeon Gold 6530 CPU.

[3]The categories cover major requirements, general education requirements, core general education, foundational courses, graduation requirements and projects, broad versus intensive majors, general education areas, department-specific major electives, graduation-credit breakdowns, and credit recognition for double majors.

**Metrics.** We evaluate retrieval using MRR and Recall@$\{1, 5, 10, 50\}$. Generation is evaluated using three metric types: (1) lexical metrics (ROUGE-L and Token-F1), which measure surface-level overlap with reference answers; (2) matching metrics (Keyword Match and Source Match), which assess whether the answer includes the required key information and uses the correct sources; and (3) the LLM-based metric (Grounded Correctness), which evaluates whether the answer is both correct and grounded in appropriate sources.[4] Further details are provided in Appendix D.

## 4 Results and Discussion

**Comparison with RAG Baselines.** Table 1 compares proFILL with RAG baselines using the same generator but different retrieval strategies. Hybrid combines BM25 and dense retrieval, HyDE (Gao et al., 2023a) augments the query with LLM-generated context, and Reranker applies a cross-encoder to refine the retrieved results.[5]

Overall, proFILL achieves the strongest retrieval and generation performance. In the *deployment* setting, it outperforms all baselines in MRR, including those given complete profiles under the *oracle* setting ($0.593$ vs. $0.475$). It also shows clear gains in lexical and matching metrics, including Source Match ($0.749$ vs. $0.412$), and achieves the highest Grounded Correctness in both settings. Interestingly, adding a profile-blind reranker to proFILL degrades performance, indicating that reranking without profile information can be harmful once profile applicability has already been incorporated into retrieval. proFILL also remains efficient, requiring $6.2$ seconds per query, compared with $6.9$ seconds for HyDE. The $1.8$-second gap from oracle proFILL ($4.4$ seconds) reflects the additional cost of on-demand adaptive profiling.

**Human Preference Evaluation.** To evaluate user preference within hoBIT's actual target population, we conduct a blind pairwise study with 48 students from our college. Participants compare proFILL against conventional dense retrieval-based RAG on ten profile-dependent questions (Appendix E). In Figure 3, proFILL is preferred on all ten questions, achieving an aggregate non-tie win rate of $85.3\%$ (354 wins vs. 61 losses; two-

[4]We use `deepeval` (Confident AI, 2024) and report scores averaged across `Qwen2.5-32B-Instruct` and `gpt-4o-mini`.

[5]We use `BAAI/bge-reranker-v2-m3`.

| | Retrieval System | Retrieval | | | | | Generation | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MRR | R@1 | R@5 | R@10 | R@50 | ROUGE-L | Token-F1 | Keyword Match | Source Match | Grounded Correctness | s/q |
| *Deployment* | BM25 | 0.089 | 0.017 | 0.118 | 0.267 | 0.728 | 0.139 | 0.167 | 0.564 | 0.412 | 0.412 | 4.0 |
| | Dense | 0.031 | 0.011 | 0.022 | 0.049 | 0.400 | 0.149 | 0.176 | 0.548 | 0.217 | 0.292 | 4.3 |
| | Hybrid | 0.061 | 0.009 | 0.069 | 0.152 | 0.698 | 0.146 | 0.173 | 0.574 | 0.380 | 0.395 | 4.3 |
| | +reranker | 0.111 | 0.027 | 0.165 | 0.322 | 0.698 | 0.152 | 0.177 | 0.621 | 0.391 | 0.425 | 4.6 |
| | HyDE | 0.061 | 0.013 | 0.057 | 0.141 | 0.689 | 0.147 | 0.175 | 0.586 | 0.367 | 0.403 | 6.9 |
| | +reranker | 0.103 | 0.025 | 0.156 | 0.300 | 0.690 | 0.151 | 0.174 | 0.613 | 0.379 | 0.422 | 7.2 |
| | proFILL | **0.593** | **0.492** | **0.718** | **0.782** | **0.936** | **0.172** | **0.199** | **0.699** | **0.749** | **0.625** | 6.2 |
| | +reranker | 0.420 | 0.271 | 0.561 | 0.691 | **0.936** | 0.168 | 0.193 | 0.665 | 0.694 | 0.596 | 6.4 |
| *Oracle* | BM25 | 0.211 | 0.059 | 0.329 | 0.624 | 0.979 | 0.155 | 0.184 | 0.648 | 0.646 | 0.555 | 4.0 |
| | Dense | 0.475 | 0.306 | 0.693 | 0.871 | 0.994 | 0.171 | 0.200 | 0.725 | 0.723 | 0.617 | 4.3 |
| | Hybrid | 0.401 | 0.219 | 0.647 | 0.841 | 0.993 | 0.171 | 0.199 | 0.716 | 0.705 | 0.621 | 4.4 |
| | +reranker | 0.151 | 0.033 | 0.224 | 0.444 | 0.993 | 0.160 | 0.186 | 0.623 | 0.583 | 0.530 | 4.5 |
| | HyDE | 0.424 | 0.242 | 0.683 | 0.886 | 0.992 | 0.169 | 0.198 | 0.716 | 0.695 | 0.608 | 7.1 |
| | +reranker | 0.159 | 0.039 | 0.229 | 0.487 | 0.993 | 0.164 | 0.191 | 0.643 | 0.591 | 0.538 | 7.2 |
| | proFILL | **0.780** | **0.674** | **0.929** | **0.980** | **1.000** | **0.179** | **0.207** | **0.733** | **0.847** | **0.676** | 4.4 |
| | +reranker | 0.549 | 0.397 | 0.704 | 0.829 | **1.000** | 0.174 | 0.201 | 0.685 | 0.787 | 0.638 | 4.5 |

Table 1: Retrieval and generation performance. For all methods, `gpt-4o-mini` is used as the answer generator. *s/q* denotes end-to-end latency in seconds per query; absolute latency may vary with the hardware configuration and external LLM API response time.

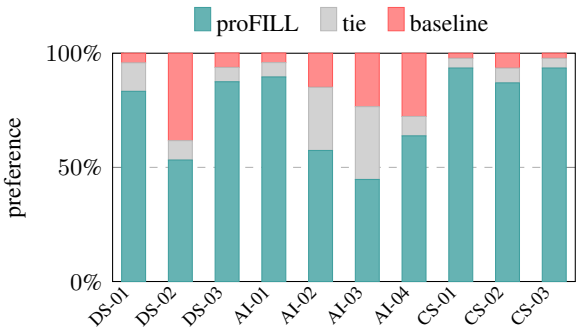


Figure 3: Human preference comparison. Results are reported as three-way vote splits. Question IDs are prefixed by the corresponding department.

sided binomial test, $p < 0.001$). Preference is highest for graduation- and credit-related questions (93–98%; DS-01, AI-01, CS-02, CS-03), whose answers strongly depend on the student's profile, and lowest for the broader question on available major courses (58%; DS-02).

**Ablation Study.** Table 2 reports ablation results for proFILL. Both profile-injection mechanisms are critical: removing either the soft prefix or hard filter substantially degrades retrieval performance. Disabling re-retrieval after evidence-driven profiling also consistently degrades performance. Notably, Recall@50 remains near ceiling without re-retrieval, indicating that its primary benefit is to promote relevant sources to higher ranks, allowing the generator to access necessary evidence with fewer chunks and use its context window more efficiently. These trends are consistent across all three dense embedders.

| Embedder | Setting | MRR | R@1 | R@5 | R@10 | R@50 |
|---|---|---|---|---|---|---|
| `openai_small` | proFILL | **0.596** | **0.495** | **0.721** | **0.784** | **0.936** |
| | w/o Hard filtering | 0.311 | 0.157 | 0.502 | 0.671 | 0.923 |
| | w/o Soft aug. | 0.247 | 0.137 | 0.337 | 0.501 | 0.893 |
| | w/o Re-retrieval | 0.580 | 0.481 | 0.702 | 0.767 | 0.934 |
| `BGE-M3` | proFILL | **0.665** | **0.579** | **0.775** | **0.796** | 0.936 |
| | w/o Hard filtering | 0.475 | 0.349 | 0.624 | 0.704 | 0.934 |
| | w/o Soft aug. | 0.347 | 0.197 | 0.517 | 0.709 | **0.938** |
| | w/o Re-retrieval | 0.644 | 0.558 | 0.750 | 0.776 | 0.934 |
| `Qwen3-Emb` | proFILL | **0.651** | **0.564** | **0.763** | **0.801** | **0.931** |
| | w/o Hard filtering | 0.425 | 0.304 | 0.574 | 0.708 | 0.926 |
| | w/o Soft aug. | 0.304 | 0.184 | 0.408 | 0.556 | 0.919 |
| | w/o Re-retrieval | 0.636 | 0.549 | 0.748 | 0.785 | 0.927 |

Table 2: Ablation study. Results are reported with three different dense embedding models for retrieval: `text-embedding-3-small` (OpenAI, 2024), BGE-M3 (Chen et al., 2024), and Qwen3-Embedding (Qwen Team, 2025).

| | Generator | *Lexical* | | *Matching* | | |
|---|---|---|---|---|---|---|
| | | ROUGE-L | Token-F1 | Kw. Match | Src. Match | GC |
| *Prop.* | `gpt-4o-mini` | 0.179 | 0.207 | 0.733 | 0.847 | 0.676 |
| *Open-weight* | `Qwen3-8B` | 0.169 | 0.197 | 0.721 | 0.879 | 0.670 |
| | `Ministral-8B` | 0.155 | 0.177 | 0.723 | 0.734 | 0.639 |
| | `LLaMA3.1-8B` | **0.204** | **0.230** | 0.665 | 0.817 | 0.649 |
| | `EXAONE3.5-7.8B` | 0.135 | 0.158 | 0.704 | **0.898** | **0.712** |
| | `Kanana1.5-8B` | 0.123 | 0.143 | **0.752** | 0.861 | 0.697 |

Table 3: Results across various LLMs, including proprietary and open-weight models. GC denotes Grounded Correctness.

**Results with Varying LLMs.** In Table 3, we evaluate proFILL with various LLM generators. We compare `gpt-4o-mini` with three multilingual open-weight models (`Qwen3-8B`, `Ministral-8B`, and `LLaMA3.1-8B`) and two Korean-specialized models (`EXAONE3.5-7.8B` and `Kanana1.5-8B`). Open-weight models remain competitive with `gpt-4o-mini`; notably, the Korean-specialized models perform strongly on matching metrics, suggesting their suitability for grounded advising in Korean-language settings.

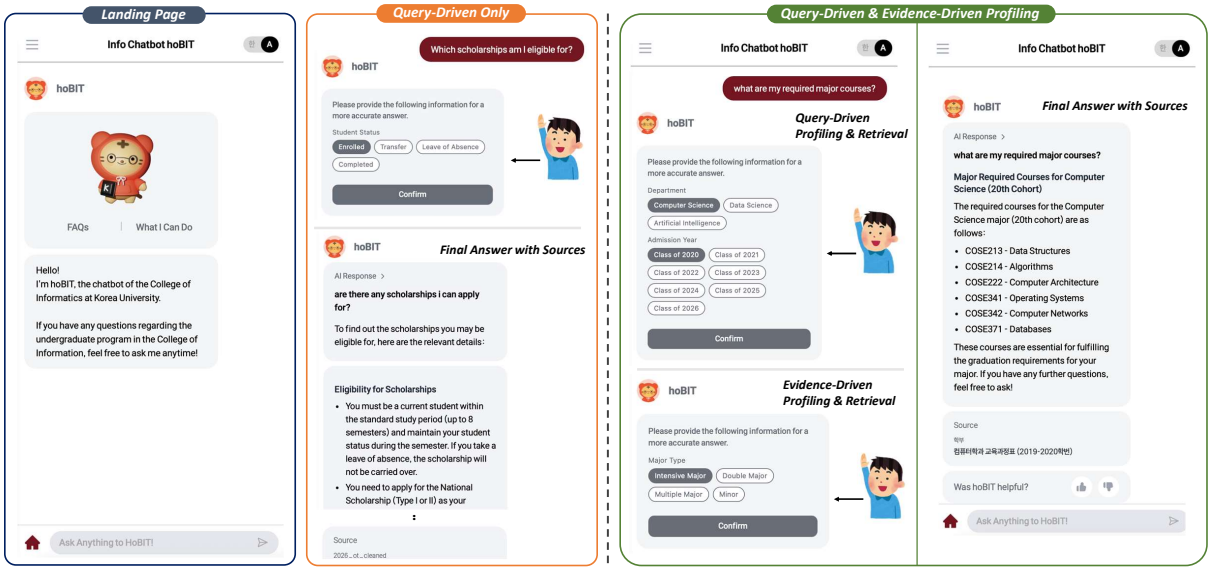


Figure 4: The web interface with examples for two queries. proFILL requests only the information needed for each query on demand and uses it to answer the user's question.

## 5 System Demonstration

Figure 4 illustrates two separate interactions with hoBIT using proFILL. For the first scholarship query, query-driven profiling obtains the student's status. As this information is sufficient to answer the query, no additional profiling is triggered, and the system directly provides an answer with supporting sources.

For the second major-course query, proFILL first collects the department and admission cohort. The initially retrieved evidence then indicates that the applicable curriculum depends on the student's major type, triggering an additional evidence-driven profiling step. After obtaining this information, proFILL re-retrieves the relevant documents and returns a profile-specific answer with supporting sources.

## 6 Related Work

**Retrieval-augmented generation.** RAG (Lewis et al., 2020) grounds generation in retrieved evidence using dense (Karpukhin et al., 2020), lexical (Robertson and Zaragoza, 2009), or hybrid retrieval (Cormack et al., 2009), and can be improved through reranking (Nogueira and Cho, 2019), query augmentation (Gao et al., 2023a), and structured indexing (Edge et al., 2024). proFILL instead conditions retrieval on an explicit user profile.

**Personalization in LLMs.** Prior work personalizes LLMs using user preferences (Choi et al., 2025; Thonet et al., 2025), interaction histories (Qin et al., 2025; Li et al., 2025; Su et al., 2025), or personas (Zerhoudi and Granitzer, 2024). In contrast, proFILL injects an explicit schema-typed profile into retrieval through soft query augmentation and hard metadata filtering, without per-user training.

**Academic-advising chatbots.** University assistants have evolved from intent-based dialogue to retrieval-grounded systems. The closest system, Marcel (Trienes et al., 2025), answers questions from university resources but does not explicitly condition retrieval on a student profile. proFILL instead uses adaptive profiling to retrieve cohort-specific evidence and tailor answers to each student.

## 7 Conclusion

We present proFILL, which transforms hoBIT from a rule-based chatbot into a profile-aware RAG system through offline profile indexing and on-demand adaptive profiling. It acquires missing information through query-driven profiling and requests additional details through evidence-driven profiling when needed. Experiments on institutional data provide practical insights into how structured profiles improve RAG for academic advising.

## Limitations

First, the corpus and benchmark are drawn from a single college of informatics. Although proFILL's overall pipeline is domain-general, its profile schema must be adapted to the curricula, poli-

cies, and advising practices of each educational institution. Second, proFILL relies on self-reported profile information without institutional verification, so incorrect inputs (e.g., an inaccurate admission year) may propagate to the retrieval results.

## Ethics Statement

**Data and Privacy.** By design, proFILL acquires only the profile attributes required for the current query, retains them only for the current session, and discards them afterward; no persistent per-user profile database is maintained. The usage logs used in this work contain no personally identifiable information and serve only to validate the query distribution—never to generate benchmark items, whose student profiles are entirely synthetic. The human preference study was voluntary and recorded only participants' department affiliations and pairwise preferences.

**Responsible Use.** hoBIT is an assistive information tool rather than an authoritative source. Every answer includes citations to institutional sources for verification, and any binding decision remains subject to the university's official regulations.

## Acknowledgments

We thank the Korea University College of Informatics and the KU Web Development Club. This work was supported by the IITP grant funded by the MSIT (IITP-2026-RS-2020-II201819), the NRF grant funded by the MSIT (RS-2026-25486220), and the MSIT under the ITRC program (IITP-2026-RS-2024-00436887), supervised by the IITP.

## References

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Youngbin Choi, Seunghyuk Cho, Minjong Lee, Moon-Jeong Park, Yesong Ko, Jungseul Ok, and Dongwoo Kim. 2025. CoPL: Collaborative preference learning for personalizing LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12875–12893.

Confident AI. 2024. DeepEval: The LLM evaluation framework. `https://github.com/confident-ai/deepeval`.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023a. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Yingqiang Gao, Fabian Winiger, Patrick Montjourides, Anastassia Shaitarova, Nianlong Gu, Simon Peng-Keller, and Gerold Schneider. 2025. SpiritRAG: A Q&A system for religion and spirituality in the united nations archive. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 26–41.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Dongsuk Jang, Ziyao Shangguan, Kyle Tegtmeyer, Anurag Gupta, Jan T. Czerminski, Sophie Chheang, and Arman Cohan. 2025. MedTutor: A retrieval-augmented LLM system for case-based medical education. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 319–353.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xintong Li, Jalend Bantupalli, Ria Dharmani, Yuwei Zhang, and Jingbo Shang. 2025. Toward multi-session personalized conversation: A large-scale dataset and hierarchical tree framework for implicit reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11493–11506.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

OpenAI. 2024. New embedding models and API updates (text-embedding-3). https://openai.com/index/new-embedding-models-and-api-updates/.

Weicong Qin, Yi Xu, Weijie Yu, Teng Shi, Chenglei Shen, Ming He, Jianping Fan, Xiao Zhang, and Jun Xu. 2025. Similarity = value? consultation value-assessment and alignment for personalized search. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9839–9852.

Qwen Team. 2025. Qwen3-Embedding. https://huggingface.co/Qwen/Qwen3-Embedding-0.6B.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Hang Su, Yun Yang, Tianyang Liu, Xin Liu, Peng Pu, and Xuesong Lu. 2025. Personalized question answering with user profile generation and compression. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4744–4763.

Thibaut Thonet, Germán Kruszewski, Jos Rozen, Pierre Erbacher, and Marc Dymetman. 2025. FaST: Feature-aware sampling and tuning for personalized preference alignment with limited data. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9341–9370.

Jan Trienes, Anastasiia Derzhanskaia, Roland Schwarzkopf, Markus Mühling, Jörg Schlötterer, and Christin Seifert. 2025. Marcel: A lightweight and open-source conversational agent for university student support. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 181–195.

Saber Zerhoudi and Michael Granitzer. 2024. PersonaRAG: Enhancing retrieval-augmented generation systems with user-centric agents. In *Information Retrieval's Role in RAG Systems (IR-RAG) Workshop at SIGIR*.

## Appendix

All system and LLM-judge prompts, the dataset, and the human evaluation questionnaire are available in our code repository (link) and on the project website (link).

## A Offline Indexing

Each cleaned document is segmented into self-contained semantic chunks of up to 800 characters using paragraph-level splitting followed by sentence-level splitting. Before indexing, each chunk is prepended with its topic, subtopic, title, and extracted keywords to improve lexical matching. Expired notices are removed, and administrator-FAQ entries are consolidated. An LLM tagger (`gpt-4o-mini`) annotates each chunk over the five predefined profile attributes. For each attribute, it assigns the value or range of students to whom the chunk applies and sets unrelated attributes to `null`. The resulting non-null fields explicitly encode the chunk's profile-dependent applicability.

Chunks are divided by update frequency into a *static* collection for stable content and a *dynamic* collection for frequently updated notices. Each collection maintains sparse and dense indices for hybrid retrieval. The sparse index uses Kiwi-based Korean tokenization with content-morpheme filtering and 32K-dimensional feature hashing for BM25 scoring. The dense index uses `text-embedding-3-small` vectors.

## B Retrieval Process

Given a query, we first translate non-Korean input into Korean. It then performs hybrid retrieval independently over the static and dynamic collections and combines their results through time-aware aggregation to construct the final retrieval results.

**Hybrid Retrieval.** Each collection is searched using both dense and BM25 retrieval, capturing overall semantic relevance and lexical term matching, respectively. The two rankings are combined using reciprocal rank fusion with $k$=60.

**Time-Aware Aggregation.** The system internally determines whether a query is time-sensitive and selects ten chunks accordingly. For time-insensitive queries, it selects seven results from the static collection and three from the dynamic collection; this allocation is reversed for time-sensitive queries. Within the dynamic collection, a recency score with a 90-day half-life favors newer notices among similarly relevant results.

## C Dataset Construction and Supplementary Results

All datasets were constructed under the guidance of our college's academic affairs office and using query logs from the deployed hoBIT service, with `GPT-4o-mini` employed for LLM-assisted generation. No personally identifiable information was collected during dataset construction.

**Profile-grounded QA.** The index contains 906 chunks segmented and profile-annotated from the collected institutional sources (§2). The main dataset is constructed solely from the static collection, and retrieval during evaluation is restricted to the same collection. It comprises 1,800 QA instances, covering all combinations of 60 student profiles, 10 profile-dependent advising categories, and three query types: formal, first-person, and verification-style. The 60 profiles combine 15 department–admission-year cohorts across CS, DS, and AI with four grade levels. Admission cohorts are further grouped into eight curriculum-revision periods, which determine the applicable curriculum for each profile.

Each instance includes a deterministic gold source and expected keywords validated against the index. Profile information is provided only through the session profile and is excluded from the query text, isolating the effect of profile-aware retrieval. The advising categories were curated from the indexed materials and cross-checked against 797 query anchors distilled from 3,058 historical hoBIT logs, covering 14 of the 15 observed log categories.

**Intent Routing.** We construct an intent-routing dataset of 1,600 queries covering five intents: *greeting*, *ability*, *faq*, *smalltalk* and *retrieval*. The dataset is assembled from three sources. First, we manually label 22 queries from real hoBIT service logs to obtain seed examples for the four non-retrieval intents. Second, using these seeds, we generate 378 additional queries using the LLM, resulting in 100 queries for each non-retrieval intent and 400 queries in total. Third, we reuse the 1,200 queries from the open-ended advising dataset as *retrieval* queries, ensuring that the routing and downstream RAG evaluations follow the same domain distribution. Table 4 shows that the proposed frame-

| Intent class | Precision | Recall | F1 | #Queries |
|---|---|---|---|---|
| greeting | 1.000 | 0.630 | 0.773 | 100 |
| ability | 0.842 | 0.960 | 0.897 | 100 |
| faq | 0.971 | 1.000 | 0.985 | 100 |
| smalltalk | 0.704 | 1.000 | 0.826 | 100 |
| retrieval | 1.000 | 0.981 | 0.990 | 1,200 |

Table 4: Intent classification results.

| Category | Top-3 Precision | Completeness | #Queries |
|---|---|---|---|
| Academic status | 0.713 | 0.973 | 100 |
| Facilities/dining | 0.683 | 0.931 | 100 |
| Academic operations | 0.670 | 0.935 | 100 |
| Scholarships | 0.643 | 0.933 | 100 |
| Space/facility | 0.633 | 0.944 | 100 |
| Enrollment/tuition | 0.593 | 0.921 | 100 |
| Clubs/council | 0.610 | 0.909 | 100 |
| Course registration | 0.593 | 0.926 | 100 |
| Student services | 0.567 | 0.935 | 100 |
| Graduate/research | 0.610 | 0.907 | 100 |
| Notices/schedule | 0.403 | 0.871 | 100 |
| Career/internship | 0.279 | 0.880 | 99 |
| **Overall** | 0.584 | 0.922 | 1,199 |

Table 5: Open-ended advising results by category.

work accurately identifies query intents, achieving an F1 score of 0.990 for *retrieval* queries.

**Open-ended Advising.** We additionally evaluate open-ended advising questions spanning 12 academic and student-life categories. Since these questions have no deterministic reference answers, we use LLM judges to evaluate two metrics (Table 5). *Top-3 Precision* measures the proportion of the top-3 retrieved chunks judged relevant to the question. *Answer Completeness*, implemented with `deepeval`, measures how fully the answer resolves the question given the retrieved context. Answers that fully use the available evidence score highly, whereas evasive or hallucinated responses score poorly. An explicit "information not found" response is rewarded when the retrieved context genuinely lacks the answer.

## D Evaluation Metrics and LLM Judges

**Lexical and Matching Metrics.** *ROUGE-L* and *Token-F1* measure overlap with reference answers after both texts are tokenized into Korean content morphemes using Kiwi, reducing penalties from inflectional variation. *Keyword Match* measures the proportion of expected content keywords included in the answer. *Source Match* evaluates attribution over the static and dynamic collections: a cited gold source scores 1, one retrieved but not cited scores 0.5, and a missing source scores 0, thereby rewarding grounded citation beyond retrieval alone.

| ID | Query Topic | Win | Tie | Lose | Win% |
|---|---|---|---|---|---|
| DS-01 | Graduation requirements | 40 | 6 | 2 | 95 |
| DS-02 | Available major courses | 25 | 4 | 18 | 58 |
| DS-03 | Required major courses | 42 | 3 | 3 | 93 |
| AI-01 | Graduation requirements | 43 | 3 | 2 | 96 |
| AI-02 | GE-required courses | 27 | 13 | 7 | 79 |
| AI-03 | Foundational courses | 21 | 15 | 11 | 66 |
| AI-04 | Required major courses | 30 | 4 | 13 | 70 |
| CS-01 | Recommend major courses | 43 | 2 | 1 | 98 |
| CS-02 | Graduation requirements | 40 | 3 | 3 | 93 |
| CS-03 | Credits to graduate | 43 | 2 | 1 | 98 |
| Total | | 354 | 55 | 61 | **85.3** |

Table 6: Results of the per-question human preference evaluation.

**LLM-based Metrics.** We evaluate generation quality using *Grounded Correctness* (GC), which combines LLM-judged *Answer Correctness* (AC) and *Source Match* (SM):

$$\mathrm{GC} = \sqrt{\mathrm{AC} \times \mathrm{SM}}.$$

The geometric mean rewards answers only when they are both correct and grounded in the appropriate evidence. *Answer Correctness* is implemented with `deepeval`: questions are classified as *positive-verification*, *negative-verification*, or *general*; verification questions evaluate only the requested yes/no decision, while general questions assess coverage of the expected key items. Scores are averaged across two independent judges on a 600-case subset stratified by the 10 categories and 3 phrasing types (60 per category).

## E Human Evaluation Details

The blind pairwise study involved 48 participants from our college: 35 from Computer Science, 7 from Data Science, 2 from Artificial Intelligence, and 4 graduate students or students from other programs. Among them, 37 completed the Korean questionnaire and 11 completed the English version. No personally identifiable information was collected during the study. All participants evaluated the same ten questions regardless of their own department, comparing responses generated under the corresponding student profile. For each of the 10 questions, participants were shown an A/B pair comparing proFILL with dense retrieval-based RAG under the *deployment* setting. To mitigate presentation-order bias, the positions of the proFILL and baseline responses in each A/B pair were counterbalanced across participants. Table 6 presents the per-question results.