

AdvisingWise: Supporting Academic Advising in Higher Education Settings Through a Human-in-the-Loop Multi-Agent Framework

Wendan Jiang*
Tufts University
Medford, MA, United States

Shiyuan Wang*
Tufts University
Medford, MA, United States

Hiba Eltigani
Tufts University
Medford, MA, United States

Rukhshan Haroon
Tufts University
Medford, MA, United States

Abdullah Bin Faisal
Tufts University
Medford, MA, United States

Fahad Dogar
Tufts University
Medford, MA, United States

Abstract

Academic advising is critical to student success in higher education, yet high student-to-advisor ratios limit advisors’ capacity to provide timely support, particularly during peak periods. Recent advances in Large Language Models (LLMs) present opportunities to enhance the advising process. We present AdvisingWise, a multi-agent system that automates time-consuming tasks, such as information retrieval and response drafting, while preserving human oversight. AdvisingWise leverages authoritative institutional resources and adaptively prompts students about their academic backgrounds to generate reliable, personalized responses. All system responses undergo human advisor validation before delivery to students. We evaluate AdvisingWise through a mixed-methods approach: (1) expert evaluation on responses of 20 sample queries, (2) LLM-as-a-judge evaluation of the information retrieval strategy, and (3) a user study with 8 academic advisors to assess the system’s practical utility. Our evaluation shows that AdvisingWise produces accurate, personalized responses. Advisors reported increasingly positive perceptions after using AdvisingWise, as their initial concerns about reliability and personalization diminished. We conclude by discussing the implications of human-AI synergy on the practice of academic advising.

1 Introduction

Academic advising plays a vital role in promoting student success and supporting career development in higher education. Empirical studies have shown that positive advising experiences are associated with improved academic performance, greater self-perceived learning gains, and stronger academic integration among students [5, 22, 35]. However, delivering high-quality academic advising remains challenging for many institutions. The absence of professional advisors often results in an over-reliance on academic professors, who, despite their disciplinary expertise, frequently lack formal training in institutional policies, making policy interpretation time-consuming and error-prone [16, 17]. Additionally, high student-to-advisor ratios often overwhelm advisors, leading to advisors’ emotional exhaustion and diminished sense of personal accomplishment [42, 52].

Given the increasing use of generative AI in a wide range of settings (e.g., customer service, tutoring, etc), it is natural to consider

whether it can be used for academic advising. While early studies show the promise of generative AI in low-stakes educational scenarios [11, 18, 25], academic advising presents a more daunting scenario because of three key challenges. First, existing generative AI models like ChatGPT [2, 4] lack the reliability required for high-stakes advising decisions due to hallucinations [15, 19, 49]. Inaccuracies in academic advising can have serious consequences, such as delaying graduation or derailing career paths. Second, effective advising requires personalization based on each student’s unique circumstances [20, 26, 29]. General-purpose language models, lacking access to student-specific information, may produce factually correct but generic responses that fail to address individual needs. Third, as revealed in our user study, advising often involves more than information delivery—emotions and trust play an important role, making human involvement an important part of the process. While prior efforts have explored AI-enhanced advising systems [32, 44, 46], these approaches fall short in two ways: over-automating at the expense of human connection, or failing to provide reliable, personalized guidance.

We present AdvisingWise, a multi-agent system, built on a core principle: automating information-intensive tasks while preserving human oversight and connection. The system provides three key properties: reliability (ensuring that answers are accurate), personalization (they are tailored based on the student’s background), and advisor-in-the-loop (all answers undergo advisor review).

To ensure *reliability*, the system employs a three-phase architecture encompassing query processing, ReAct-style [50] information retrieval, and quality-controlled response generation. For *personalization*, it maintains student profiles based on conversation history and adaptively prompts students for additional academic background when needed. To ensure *advisor-in-the-loop*, all draft responses—including detailed answers, summaries, and lists of cited sources—undergo advisor review, editing, and approval before delivery to students. To support this review process, the system provides advisor-only notes that flag any uncertainties or limitations in the draft response.

We implemented and evaluated AdvisingWise for computer science graduate advising at our institution, a private U.S. university. The prototype uses both GPT-4o and GPT-4o-mini models [2]. Users interact with the system through RocketChat [40], an open-source communication platform. Our evaluation uses a mixed-methods approach. To evaluate response correctness, we conducted expert evaluation on 20 queries sampled from 100 benchmark questions. For

*Both authors contributed equally to this research.

the 19 queries the expert could evaluate, 84.2% (16 of 19) achieved high accuracy scores. Lower-scoring responses resulted from the system’s conservative approach: it declined to answer when uncertain. We used LLM-as-a-judge pairwise comparison [53] to evaluate our information retrieval mechanism against a standard RAG baseline [31]. The results show that our ReAct-style retrieval system outperformed the baseline, with notable strength on inference-based queries, where only general or scattered information was available.

We conducted a user study with eight academic advisors examining their attitudes toward AI integration before and after using AdvisingWise. Direct engagement with the system positively shifted perceptions: advisors who initially expressed reservations about AI’s ability to deliver accurate and personalized guidance demonstrated increased willingness to adopt the system. Concerns diminished as they observed AdvisingWise successfully handle complex, context-dependent queries with accurate and nuanced responses.

In summary, the main contributions of this paper are:

- (1) **Human-Centered System Design** A human-in-the-loop system design that maintains both students and advisors in the workflow. Students provide context for personalization, and advisors validate outputs before delivery, preserving essential human judgment and connection.
- (2) **Multi-Agent Architecture** A three-phase multi-agent architecture that delivers correct and contextually appropriate responses.
- (3) **Mixed-Methods Evaluation** Mixed-methods evaluation comprising expert assessment, LLM-as-a-judge comparison, and user study with 8 advisors, demonstrating high effectiveness.

2 Related Work

This work builds on a growing body of research exploring the application of AI to academic advising and the development of human-in-the-loop frameworks that ensure effective oversight and collaboration between humans and AI.

2.1 Academic Advising in CS Education

Academic advising is a critical part of higher education. It provides students with support to navigate their educational program in a manner that helps them achieve their goals. It also serves as a first line of defense in majors like computer science, where students face a higher risk of dropping out [33, 41]. However, providing quality advising becomes increasingly challenging when advisors at research universities face mounting pressures from expanding student populations while managing competing demands for research productivity and teaching excellence [9, 52].

Zhang documented these workload challenges through interviews with engineering faculty advisors, revealing that advisors struggle to balance personalized guidance with limited time resources, often resulting in rushed meetings that fail to address students’ complex needs. In addition, hiring full-time academic advisors may not be feasible given resource limitations [3]. Beyond workload pressures and resource constraints, advisors must

also keep track of a plethora of resources and information to provide accurate planning, which may require additional training and awareness whenever new requirements are introduced [7, 52].

Méndez et al. observed that advisors must synthesize information about students’ academic histories, personal circumstances, and career aspirations while considering program requirements, prerequisite chains, and course availability—all within brief advising sessions. This complexity is further illustrated in queries requiring multi-source reasoning, such as determining whether “Build LLM Agent” counts toward graduation requirements—a question that cannot be answered from a single document but requires synthesizing course listings, requirement definitions, and potential naming variations across different systems [33].

These challenges highlight the need for tools that can support advisors by integrating distributed resources and handling complex queries. To that end, we developed an advising chatbot (AdvisingWise) to answer students’ questions and provide advisors with timely access to up-to-date information.

2.2 AI and LLMs in Academic Advising

Traditional machine learning algorithms have long been used to address the challenges of academic advising. For example, Méndez et al. examined a course performance prediction tool to support advisors in planning and making recommendations. Advisors responded positively but raised concerns about accuracy and interface overload [33]. These limitations motivated researchers to explore more flexible AI approaches, particularly large language models (LLMs) known for their human-like text generation and advanced text processing.

Early explorations relied on general-purpose chat applications such as ChatGPT¹, typically without specialized advising interfaces. Akiba and Fraboni evaluated ChatGPT on a small set of advising questions, finding strong performance on general queries but weaker accuracy on individualized ones [3]. Similarly, Lekan and Pardos investigated its use for undergraduate major recommendations; while advisors rated its suggestions as helpful, they agreed with them only 33% of the time [30]. These studies suggest that general-purpose LLMs can assist advising but cannot yet replicate expert judgment.

Other research has developed dedicated systems, but it has also revealed further gaps. ARGObot employed multiple tools (retrieval-augmented generation, search, email) to improve accuracy, yet its functions were fragmented across different interfaces such as Gmail [44]. Ma et al. proposed a recommender system that generated degree plans from student input but lacked advisor involvement in refining recommendations [32]. Seahawk required students to upload documents for course recommendations [46], while Advisely emphasized analytics of student interactions over direct advising [1]. Across these systems, limitations included missing user evaluation, over-automation, and insufficient integration of advisors’ expertise.

While prior work shows that LLMs can reduce academic staff workload [37], advising is a complex and personalized process that extends beyond accurate information to include student well-being and dropout prevention [9]. Our approach, AdvisingWise, seeks to

¹<https://openai.com/index/chatgpt/>

strike this balance by reducing workload while ensuring human advisors remain central to the interaction.

2.3 Human-in-the-Loop and Oversight Systems

The design of effective human oversight mechanisms for AI-generated content remains a critical challenge. Prior work has examined paradigms such as conditional delegation, where humans predefine rules for when AI output is to be trusted [27], and cognitive forcing functions, which deliberately compel analytic engagement with AI suggestions [8]. These studies demonstrate that even non-expert users can identify contexts in which AI decisions are reliable, and that human-AI teams often outperform either working alone or independently.

Beyond static oversight, hybrid approaches integrate human feedback into AI systems for continuous improvement. Jakubik et al. demonstrated models that learn from human review decisions to handle previously unseen cases, reducing review effort by up to 73% while maintaining accuracy. However, these gains must be weighed against the costs of misclassifications and ongoing human supervision [21]. Utilization of these systems raises concerns about user agency and the explainability of AI decisions [39].

Interface design is another crucial factor. Research consistently finds that user acceptance depends on systems being simple, familiar, and easy to integrate into existing workflows [6, 33, 39]. Dang et al. adopted the classic idea of margin annotations to give users access to evolving AI-generated summaries during writing, but they observed tensions in how users interpreted different interface components [13]. Such findings underscore that usability can make or break oversight systems.

Building on these insights, AdvisingWise delivers suggested responses and document citations through a unified interface, preserving advisor autonomy while reducing workload. Via a chatbot interface, advisors can review and edit AI-generated drafts before sharing them with students, ensuring human oversight while benefiting from the efficiency of AI support.

2.4 Multi-Agent LLM-based Systems

Academic advising requires a nuanced understanding of students' needs and circumstances [52]. Developing AI systems capable of context-aware generation in such scenarios demands deeper reasoning and efficient handling of ambiguity [14, 36]. Recently, LLM-based systems equipped with mechanisms for storing intermediate states and leveraging external tools have demonstrated strong problem-solving capabilities across multiple domains [23, 43, 47, 51]. These LLM agents can interact, reason, and autonomously decide on subsequent actions to accomplish a given task.

However, for complex tasks, greater efficiency is often realized through multi-agent systems, where multiple specialized LLM agents collaborate towards a shared goal. In such systems, each agent is assigned a specific role with tailored prompts, available actions, and dedicated memory. By reasoning, planning, executing, and delegating sub-tasks among one another, these agents collectively achieve outcomes that surpass the capabilities of individual agents [14, 36].

AI researchers and practitioners have developed new frameworks (e.g., LangGraph [28]) and design methods (e.g., ReAct [50])

to operationalize the implementation of such systems [10, 14]. Nevertheless, designing multi-agent systems requires a careful balance between what decisions are delegated to agents and the degree of human oversight needed to ensure reliable and safe behavior [14, 34, 36].

In designing AdvisingWise, we aimed to strike this balance: granting the system flexibility and tools to handle ambiguity and complex queries, while incorporating appropriate human involvement to prevent harmful consequences.

3 AdvisingWise

The system's design incorporates three core principles: (1) automating time-consuming information retrieval and drafting tasks, (2) tailoring AI-generated responses to student backgrounds, and (3) ensuring advisors remain engaged through review of all system outputs.

Achieving these principles requires careful architectural design. While automation can reduce workload, it depends critically on accuracy. Research has shown that LLMs like GPT-4o tend to generate incorrect answers rather than acknowledging uncertainty when context is insufficient [24]. To address this challenge, as shown in Fig. 1, we designed a three-phase multi-agent system—query processing, information collection, and response generation—with particular emphasis on robust information retrieval. The information collection phase implements a ReAct-style approach [50] with two specialized agents working iteratively: a reasoning agent identifies what information is needed, and an action agent retrieves it through context-dependent actions such as searching the knowledge base, web searching, or prompting students for additional information.

For personalization, the query preprocessing phase extracts academic information from student queries and updates individual student profiles. This background information guides both current response generation and future query handling, enabling contextually appropriate responses tailored to each student's situation.

To preserve advisor engagement, the system positions advisors as essential reviewers rather than replacing them. Advisors can approve drafts with one click or edit as needed, maintaining their oversight while benefiting from automated assistance.

3.1 User Interaction with AdvisingWise

AdvisingWise acts as an intelligent proxy that routes communication between students and advisors. Students message AdvisingWise with questions, which triggers notifications to advisors with AI-generated draft responses. After advisors approve or edit the drafts, AdvisingWise forwards the final answers to students. To support natural and transparent communication, we adopted Rocket.Chat [40], an open-source messaging platform, as the front-end interface. This setup allows students and advisors to interact with AdvisingWise in a familiar chat-based environment.

3.1.1 Advisor-in-the-loop. As shown in Fig. 2, when a student submits a question, AdvisingWise processes it and sends a message notification to the designated advisor. The notification shows the student's name and question. To keep the interface organized, the AI-generated answer appears as a "thread"—a collapsible, nested conversation attached to the notification message. Advisors click a

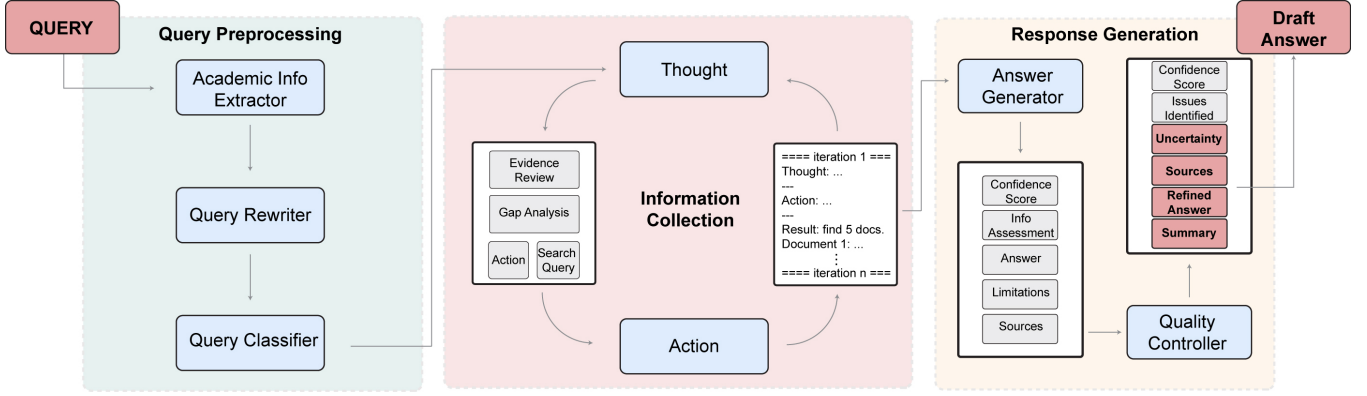


Figure 1: Three-phase multi-agent architecture. Phase 1 (Query Preprocessing): Academic Info Extractor, Query Rewriter, and Query Classifier—if classified as off-topic, the query bypasses Phases 2-3. Phase 2 (Information Collection): Thought and Action agents collaborate iteratively (up to 4 cycles) to gather information from institutional knowledge base, course database, web sources, or student prompts. Phase 3 (Response Generation): Answer Generator and Quality Controller produce draft outputs. Red-highlighted components (detailed answer, summary, cited sources, and advisor notes) appear in the final draft response delivered to advisors.

“View Thread” button to expand and review the draft answer. The question displayed in the advisor notification is reformulated by AdvisingWise to integrate conversation context.

The draft answer includes four components: (1) a response text, (2) a summary line, (3) a list of sources, and (4) an advisor-specific note when applicable. The response text provides a detailed answer to the student’s question, highlighting important information. The summary line provides a short summary of the answer for quick scanning. The sources list identifies all documents used to draft the answer. For institutional documents (e.g., handbooks, websites), it displays the document name along with relevant page numbers or web links to facilitate verification. When the draft answer draws on student background information from conversation history, the system displays a timestamped reference such as “Student Academic Profile (Last Updated: [date])”, allowing advisors to assess whether the information is current and follow up with the student if necessary. Each answer may include an advisor-only note that appears when the system cannot fully answer the student’s question (see Fig. 3). The note identifies what information is missing from available institutional resources, allowing advisors to locate which parts of the response need further revision. Advisors may review, edit, or approve AI-generated answers before sharing them with students. A “Copy to Chat” button facilitates this process by transferring the draft answer directly into the chatbox.

3.1.2 Student-in-the-loop. Advising responses depend on individual students’ academic profiles, as the same question may have different answers depending on factors like major, completed courses, or degree requirements. For instance, a course that fulfills a degree requirement for Student A may not apply the same way for Student B. AdvisingWise collects academic information directly from students through conversation, mirroring how human advisors gather context before offering guidance. Upon receiving a question, the system determines if additional student information is required and requests it through follow-up questions before generating a

draft answer. This approach enables personalized responses while minimizing back-and-forth between advisors and students. Figure 4 illustrates this interaction flow, showing how AdvisingWise prompts students for additional information when needed.

3.2 Backend Design

Queries are processed through a three-phase multi-agent system implemented in LangGraph [28], as shown in Fig. 1. The system employs a tiered model strategy: agents in the query preprocessing and information collection phases use GPT-4o-mini for cost efficiency, while the “Answer Generator” and “Quality Controller” in the response generation phase use GPT-4o to maximize accuracy. This approach balances computational cost with output quality across the workflow.

3.2.1 Query Preprocessing. This phase consists of three agents that work together to (1) ensure queries are self-contained and relevant to academic advising, and (2) extract student information to enable personalized responses. For instance, a student query like “I have taken CS101, what other core courses must I take?” contains valuable information (CS101 completion) embedded within the question itself. By extracting and storing this detail, AdvisingWise can later provide personalized responses to questions like “Can I take CS102?” (which has CS101 as a prerequisite) without needing to asking the student’s CS101 completion again.

The “Academic Info Extractor Agent” implements this functionality by extracting academic details from each student query without modifying the original text and storing them as key-value pairs in the database. Each student has an individual profile that is updated as new information is shared. After that, the “Query Rewriter Agent” converts queries into self-contained versions by integrating necessary context from the conversation history. Finally, the “Query Classifier Agent” identifies and filters off-topic queries, ensuring AdvisingWise handles only advising questions. Off-topic queries do not proceed to subsequent phases.

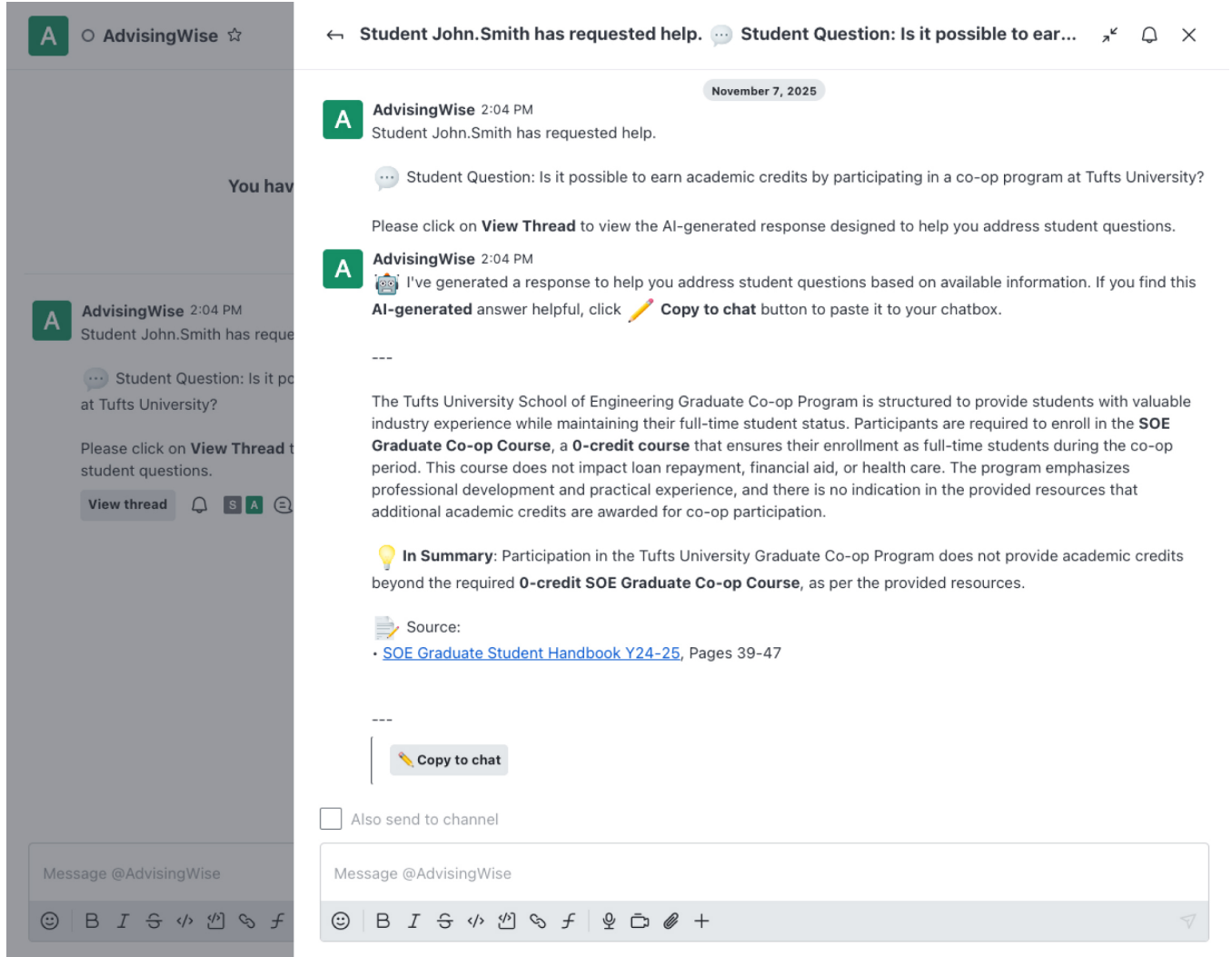


Figure 2: Advisor interface: (left) notification of student question, (right) AI-generated draft with response, summary, sources. Student names are pseudonyms.

3.2.2 Information Collection. The information collection phase follows a ReAct-style workflow [50]. In the implementation, two agents alternate between reasoning about what information is needed and taking actions to retrieve it. The “Thought Agent” analyzes the query and any previously gathered information to identify what additional data is needed, then selects an appropriate retrieval action (e.g., querying the institutional knowledge base, conducting web searches, or requesting information from students). The “Action Agent” executes this action and retrieves the requested data. The retrieved information is appended to the existing reasoning traces, forming a Chain-of-Thought (CoT) [48] that is passed to the “Thought Agent” for continued reasoning. This reasoning-action cycle repeats for up to four iterations or until the “Thought Agent” determines that all identified information gaps have been addressed.

Action selection is context-dependent and guided by advising-specific heuristics: prioritizing institutional sources for policy-related

queries and requesting information from students for personalization. To establish foundational context, the “Thought Agent” always starts with `search_knowledge_base` to retrieve relevant institutional information. In subsequent iterations, it may select `search_course_db` if the knowledge base lacks sufficient course-specific details, `search_web` when internal resources are insufficient or subjective information, such as course workload, is needed, or `request_student_info` when individualized student data, such as completed courses or expected graduation timeline, is required. Finally, when sufficient information has been gathered, “Thought Agent” selects `provide_answer` to conclude the collection phase. For actions that involve information retrieval, the “Thought Agent” generates a *search query*, which serves as the input for the “Action Agent” to execute the corresponding search. Similarly, when the action requires collecting additional academic background from the student, the “Thought Agent” generates a question as the *search*

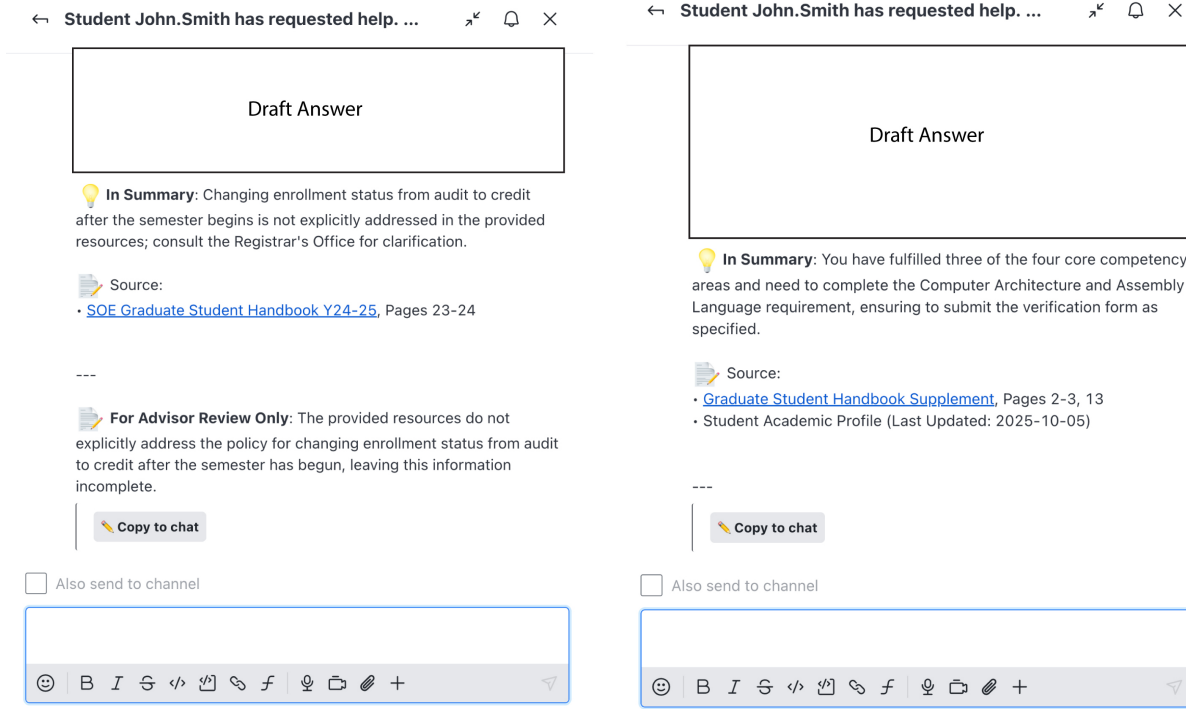


Figure 3: Example draft responses demonstrating key features: (left) advisor-only notes flag uncertainty, (right) personalized response based on student academic profile, which is explicitly cited as a source

query, which the “Action Agent” presents to the student. Only the search_web action involves the use of LLMs, while all other actions focus purely on information retrieval and consolidation.

We now detail how the “Action Agent” handles each specific action chosen by the “Thought Agent”:

- (1) search_knowledge_base: Queries an internal RAG system [31] where institutional documents are stored in a MongoDB vector store with AzureOpenAIEmbeddings. It retrieves document chunks most relevant to the Thought Agent’s query based on similarity scores. Each chunk serves as factual evidence, ensuring responses are grounded in authoritative institutional documents.
- (2) search_course_db: Queries the course database by exact course name or code to retrieve structured course information.
- (3) search_web: Searches external sources via Tavily [45] when internal resources provide insufficient information about course content or subjective experiences (e.g., workload, student experiences). Retrieved content is validated through GPT-4o [2] to ensure institution-specificity and query relevance, filtering out generic or cross-institution information.
- (4) request_student_info: Uses LangGraph’s human-in-the-loop functionality to pause the workflow and prompt students directly for personalized information, resuming automatically once they respond.
- (5) provide_answer: Proceeds to the Response Generation phase.

3.2.3 Response Generation. In this phase, the “Answer Generator Agent” receives the student’s academic profile and information from the previous phase. The agent bases its responses exclusively on this provided information without making unsupported claims or assumptions. Each response includes five structured components: (1) information completeness assessment, (2) confidence score, (3) detailed evidence-based answer, (4) source references, and (5) identified limitations or gaps. Responses use professional formatting (bold text for emphasis, structured paragraphs, bullet points) to ensure clarity and readability. The “Quality Controller Agent” validates generated responses to prevent hallucinations by verifying factual accuracy against provided documents, ensuring proper source citations, identifying information gaps, and confirming adherence to academic advising standards (clear language, professional tone). Responses failing validation are flagged for revision.

4 Experiments

We adopted a mixed-method approach to evaluate AdvisingWise. First, we assessed the accuracy of its responses through expert evaluation §4.1. Next, we used large language models (LLMs) to benchmark the responses against baselines using tested techniques §4.2. Finally, we conducted a user-centered evaluation, gathering academic advisors’ perspectives through system use and interview questions during a lab study §4.3.

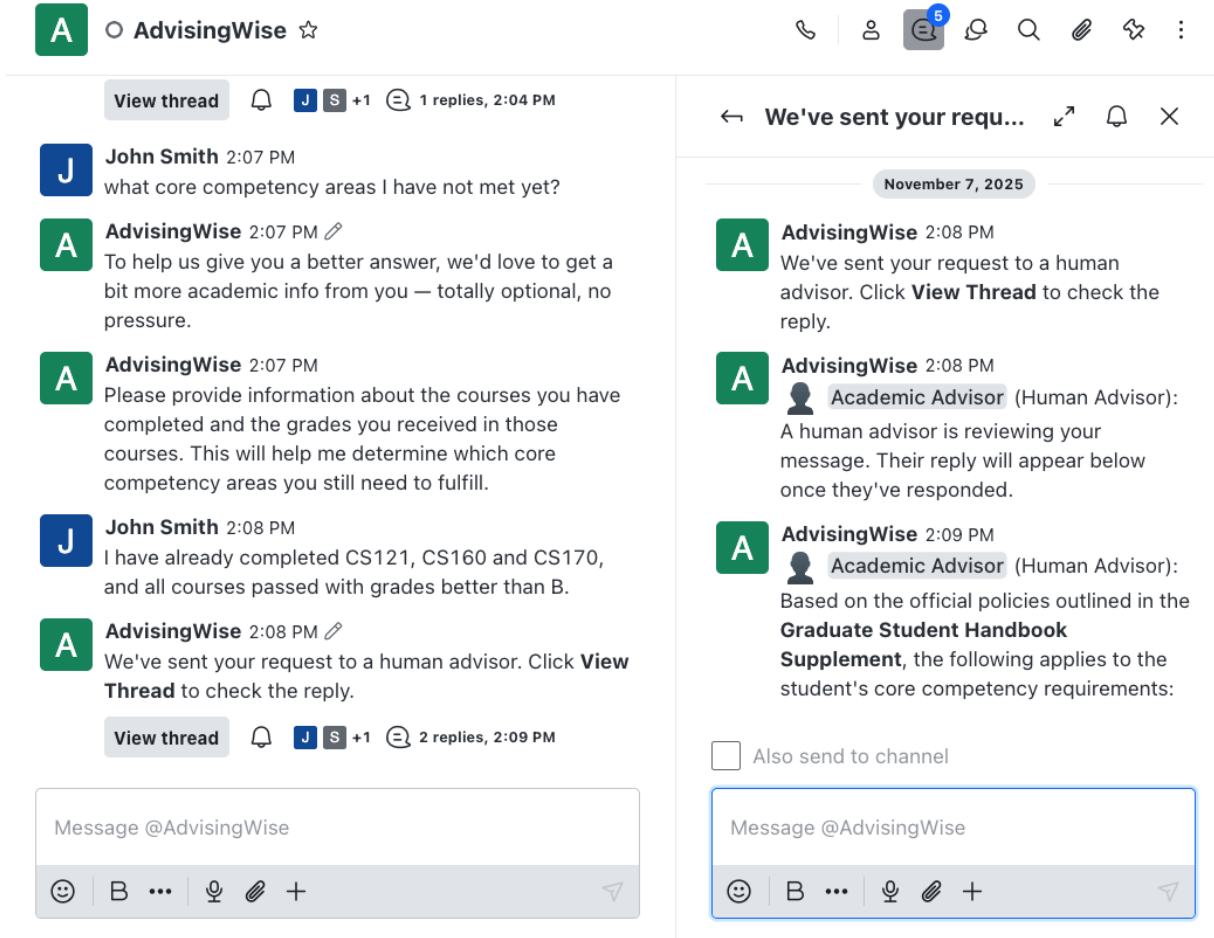


Figure 4: Student interaction flow: AdvisingWise prompts for academic background before generating personalized draft for advisor review.

4.1 Expert Validation

To evaluate AdvisingWise, we synthesized 100 queries using Anthropic’s Claude Sonnet 3.7 [4]. The queries were designed to be self-contained (i.e., not requiring additional student information) and to reflect frequently asked topics identified by experts (details in Appendix A). They addressed three common scenarios based on information availability in the institution’s handbook: (1) specific information about the topic is available (Handbook-Explicit), (2) only general or scattered information is available (Handbook-Implicit), and (3) no information is available (Handbook-Unavailable). These scenarios assess AdvisingWise’s ability to retrieve, synthesize, and interpret information from varying levels of completeness.

We used an automated Python script to process each query sequentially through the backend system. Each query was treated as originating from a distinct user session to ensure independent system responses. To create a manageable set for expert validation, we randomly sampled 20 queries using a stratified, weighted sampling strategy [12] based on topic frequency and relevance to

core advising decisions. Categories were grouped into three priority levels, determined by our university context and expert input. Within each category, questions were further weighted by information availability, with Handbook-Explicit questions assigned the highest weight and Handbook-Unavailable questions the lowest. A query’s final weight was the combination of its category priority and question type, and sampling was conducted probabilistically according to these weights.

This strategy deliberately prioritized frequently asked topics and questions with available information over comprehensive category coverage, thereby focusing expert effort on queries that AdvisingWise is most likely to answer and enabling a more targeted evaluation. Table 1 presents the sample details.

An academic advising expert evaluated the accuracy of the 20 sampled responses using a five-point scale as listed in table 2. If the expert was uncertain about a response’s accuracy, they could abstain from rating by selecting “I don’t know.”

Table 1: Queries Sample Distribution by Category and Question Type

Category Description	Handbook-Explicit	Handbook-Implicit	Handbook-Unavailable	Total by Category
Co-op	3	1	0	4
Core Competency	2	1	1	4
Degree Completion	4	0	0	4
Important Dates	1	1	1	3
Credit Transfer	2	3	0	5
Total by Type	12	6	2	20

Table 2: Rating Scale for Assessing Response Accuracy

Rating	Description
1 - Not correct at all	The response is mostly or entirely wrong, misleading, or contradicts the handbook.
2 - Mostly incorrect	Contains some relevant information, but major inaccuracies or omissions make it unreliable.
3 - Somewhat correct	The main idea is generally correct, but minor factual errors are present. The response would require clarification or correction before being shared with students.
4 - Mostly correct	Accurate overall, with only minor details that could be better aligned with the handbook.
5 - Extremely accurate	Fully consistent with the handbook, clearly stated, and requiring no corrections.

4.2 LLM-as-a-Judge Evaluation

A key challenge in evaluating our system is disentangling whether response accuracy stems from our ReAct-style information retrieval architecture (§3.2.2) or from the inherent capabilities of the underlying language model—i.e., whether the LLM could achieve comparable accuracy with simpler retrieval methods. To address this, we employed the LLM-as-a-Judge framework [53] to compare our ReAct-style approach against a standard Retrieval-Augmented Generation (RAG) baseline [31], using the expert-validated responses (§4.1) as reference answers.

Using our 20-query sample, we retrieved relevant information once with AdvisingWise’s “Information Collection” system and once with the RAG-only system (one-time retrieval of top-k relevant chunks). In both cases, the system only differ in the retrieval strategy. The retrieved information was then fed to the same LLM (GPT-4o), using the same instructions to generate responses. We then conducted a reference-guided, pairwise comparison of outputs using GPT-4o as the judge. For each query, the judge received: (1) the original query, (2) the expert-validated reference answer, (3) Response A (from one system), and (4) Response B (from the other system). The judge was not informed which system produced which response, and to mitigate position bias, each comparison was conducted twice with the response order swapped. The judge evaluated responses relative to the reference answer along two dimensions: correctness and helpfulness.

4.3 User Study

In this study, we aimed to examine academic advising workflows, assess advisors’ attitudes toward LLM integration, collect feedback on AI-generated responses, and evaluate the impact of AdvisingWise on participants’ perceptions and adoption of LLM-based tools. The study employed a semi-structured interview design to gather both quantitative data and qualitative insights. Each session lasted approximately 60 minutes. All study procedures followed the guidelines for human subjects research and were approved by our university’s Institutional Review Board (IRB).

4.3.1 Context. Our study site employs a combined advising model that incorporates two roles: staff advisors, who specialize in university-wide policies, and faculty advisors, who focus on career development and course recommendations tailored to students’ academic interests. Each student is assigned a faculty advisor in the student information system, and this assignment remains consistent throughout their time in the program. Prior to each semester’s course registration, students must meet with either their faculty advisor or a staff advisor—through individual meetings, group advising sessions, or email—to have their advising hold lifted.

4.3.2 Participants. Advisors from the CS department were recruited for the study, with eight participants completing it. All participants currently serve in academic advising roles and have direct experience with student advising processes. The pool included four staff advisors and four faculty advisors, representing the two primary advising roles within the program. Staff advisors primarily provide guidance on logistical matters and degree requirements, while faculty advisors focus on supporting students’ career development and offering course recommendations aligned with their academic interests.

4.3.3 Procedure. The study consisted of three parts, with both audio and screen interactions recorded for further analysis. The first part (10–15 minutes) of the interview captured participants’ advising practices and baseline attitudes toward AI. It began by exploring advising workflows and common challenges, followed by participants’ familiarity with and perceptions of generative AI, including prior experience with tools such as ChatGPT, perceived usefulness, and concerns about such tools in advising.

The second part involved hands-on interactions with AdvisingWise across four representative scenarios, each corresponding to a distinct question type based on information availability: Handbook-Explicit, Handbook-Implicit, and Handbook-Unavailable—along with an additional type requiring student-specific information (Student-Context Dependent; see Table 3). After a brief introduction to AdvisingWise, participants engaged with each scenario by reviewing student questions and AdvisingWise responses, with the option to edit or approve those responses. Feedback was collected for each response, and follow-up surveys captured quantitative ratings of perceived accuracy, style, and willingness for delegation. For scenarios Q2 and Q3, additional feedback addressed AdvisingWise’s acknowledgment of uncertainty, while for Q4, participants evaluated the adequacy of the student information gathered.

The final part of the interview focused on broader reflections regarding the integration of AI in academic advising. Participants provided overall assessments of AdvisingWise, highlighting the

Table 3: Details of the four question types used in the user study and their corresponding categories.

Question ID	Question	Category
Q1	Do students receive academic credit for their co-op experience in the Master of Science program at [school name]?	Handbook-Explicit
Q2	What should I do next if I receive a C+ in my Networks course?	Handbook-Implicit
Q3	Can I switch my enrollment status for a course from audit to credit after the semester has already started?	Handbook-Unavailable
Q4	What core competency areas I have not met yet?	Student-Context-Dependent

features they found most valuable and evaluating aspects of its responses that inspired either confidence or concern. This stage also explored how direct interaction with AdvisingWise influenced participants’ initial perceptions of AI in advising, and whether their concerns were mitigated or reinforced. The discussion concluded with consideration of practical implementation issues, including potential enhancements or additional functionalities to increase AdvisingWise’s utility.

4.3.4 Data Analysis. We conducted descriptive statistical analyses on the quantitative data to examine participants’ attitudes toward the use of AI and any changes following their interaction with AdvisingWise. We also summarized participants’ ratings of response accuracy, style, and willingness to delegate tasks to the system.

For the qualitative analysis, audio recordings were transcribed and manually verified for accuracy. We followed an approach similar to Park et al.[38]. Two members of the research team first reviewed two randomly selected interview transcripts (25% of participants) to develop an initial understanding of participants’ perspectives through open coding. Each coder independently performed line-by-line analysis of half of this subset ($n = 1$) to generate initial thematic codes, after which they met to compare observations and collaboratively develop a shared coding framework. Both researchers then re-analyzed the initial transcripts using the framework to ensure consistency. The remaining transcripts were evenly divided, with each researcher coding half. The researchers then met to discuss and reconcile the codes for these remaining transcripts to ensure consistency in the analysis. All themes were generated inductively from the data without predetermined categories, allowing participants’ experiences and perspectives to drive the analysis.

5 Results

AdvisingWise was positively received by human advisors, with over 60% rating it as useful and half reporting that it eased their concerns of using AI in advising. Evaluation results showed that AdvisingWise achieved over 80% accuracy based on expert validation and performed three times better than a RAG-based system under the LLM-as-a-Judge framework.

5.1 Overall Evaluation and Changed Attitudes

Academic advisors demonstrated predominantly positive attitudes toward integrating AdvisingWise into academic advising practices.

Quantitatively, the majority of participants (5 of 8, 62.5%) assigned a usefulness rating of 4 on the assessment scale. Qualitatively, half of the advisors ($n = 4$) explicitly indicated that their pre-existing concerns regarding using generative AI had been mitigated to varying extents.

5.1.1 Perceived Usefulness and Potential Concerns of AI. The majority of participants (7 of 8, 87.5%) reported prior experience with generative AI chatbots, such as ChatGPT (Table 4). However, the level of engagement varied, with one participant (P6) characterizing their usage as “very little.” Only one participant (P1) indicated no prior exposure to generative AI, stating, “I don’t know what it is.” Some advisors agreed that from their prior experience, generative AI are helpful as “a replacement for search” (P4) to give them “an understanding of how to move forward” (P7). Most advisors rated the usefulness of AI integration low, with the majority providing ratings below or equal to 3 on a 5-point scale, indicating neutral to skeptical attitudes (Fig. 5).

Most advisors raised concerns about the accuracy of responses generated by LLMs, with P1 noting, “It’s not always so reliable.” Providing incorrect information could create problems for both students and advisors: students might act on faulty guidance, while advisors could spend significant time correcting errors. P4 emphasized that even a small error rate can greatly reduce the usefulness of generative AI in advising, stating, “I don’t care if the bot is like accurate 99% of the time, because that 1% is going to make it completely useless for me, right? Because the amount of time I would waste on that one person would usually be much more.”

In addition, advisors expressed mixed views on AI’s capabilities depending on question complexity. They were optimistic about AI’s ability to answer straightforward factual questions. As P5 explained, “for the more prescriptive stuff, assuming they are trained properly and using the proper resources, I think it’s [usefulness is] a five out of five.” (P5) However, advisors doubted AI’s ability to handle complex questions requiring nuanced judgment. As P8 noted, advising encompasses a broad array of issues, including deeply personal questions, including “what do I want to do as a career?” For such highly individualized matters, advisors worried that AI cannot adequately “figure out what directions you’re (students) heading and give you (students) good advice” (P6). This skepticism stems from the inherent diversity among students. “Basically, no two students are the same ... They’re all unique ... it’s hard to

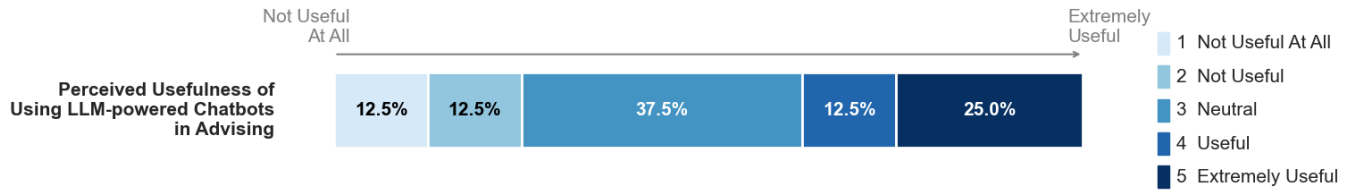


Figure 5: Perceived Usefulness of Using LLM-Powered Chatbots in Advising

even put students into similar categories with each other” (P5). While generic career advice may not be harmful, it is unlikely to provide meaningful benefit to individual students. This challenge of personalization is not limited to career advice. Advisors noted that policy questions similarly require understanding individual student contexts, as answers often depend on specific circumstances. P7 worried that students might overgeneralize AI responses: “If a student asks a question, gets a response, and they think that this is 100% the truth, and it works for every situation,” such misplaced confidence could lead students to act on advice inappropriate for their specific circumstances. In this scenario, the generic response may even bring harmful effects.

5.1.2 Easing and Persistence of Concerns with AdvisingWise. Half of the advisors (P1, P2, P5, P7) explicitly reported that their concerns regarding the integration of generative AI tools in academic advising had been alleviated, whereas others (P3, P8) indicated that their concerns persisted.

P1 and P2 initially expressed worries about the accuracy of AI-generated responses; after using AdvisingWise, both impressed by AdvisingWise’s accuracy. P2 noted, “I think it does a better job of pulling up the factual.” P1, who had no prior experience with generative AI and was initially highly concerned about accuracy, observed, “it was more accurate than I was expecting”. Although P1 acknowledged ongoing concerns regarding “incomplete or wrong answers,” they recognized the potential utility of the system, stating that they “can see where it could be helpful” and felt “a little more encouraged by that.” In contrast, P8 remained concerned about accuracy, believing that hallucinated information is inherent to LLMs, noting, “they’re fundamental.”

P5 and P7 indicated that their primary concern was the generative AI’s ability to handle nuanced and personalized situations. After using AdvisingWise, both reported that their concerns were “slightly eased” (P5). P5 appreciated that the system “seems to acknowledge that these things (personalization) exist and it asks about additional information [to handle the more personal scenarios]” Both advisors acknowledged that the level of personalization that the system performed when addressing advising questions for master’s programs is “pretty solid” (P5). Nevertheless, their concerns were only partially alleviated, as “[requirements, rules and policies in] MS in computer science is a little more standard,” whereas undergraduate advising is “a bit more personalized than just what classes have you taken,” raising doubts about whether AdvisingWise can succeed in that context (P5, P6). In contrast, P3 took a more conservative stance, emphasizing that the level of personalization provided by AdvisingWise is insufficient and that

“there’s still a need for a human to explain their advice based on the conditions.”

5.1.3 Attitudes Towards Integration in Academic Advising. Advisors are optimistic about using such a tool in their daily advising work, agreeing that the system meets the baseline requirement of accurately answering straightforward, factual questions:

“It answers a lot of pretty simple questions quite clearly and directly and gives the correct information. And a student doesn’t need to go into an advising session necessarily, and they can get the answers when they want them. It also can help people prepare better for an advising session” (P2).

P5 was particularly enthusiastic, claiming that AdvisingWise “exceeded my expectations,” and particularly expressed their optimism of seeing it deployed in the department, “I would 100% use this in its current form.” They explained that

“It could be really cool for the prescriptive stuff and it’s on its way to being super useful for The non-prescriptive stuff ... I’m so torn because I want to use it for as many things as humanly possible. It’s just right now, a lot of the things that I would use it for, it wouldn’t work-meaning a commercial one, not a dedicated one to our knowledge. I can’t ask ChatGPT anything about [the school name] ... There’s AIs being used for so many useless things, and there are so many useful things that could be useful like this ” (P5).

P3 also think the scope of the advising could be broadened up to prospective students, “I think it would help different groups earlier on. So what I mean by that is, for example, high school prospective students, in other words, students, students looking forward to certain programs at [the school name]” (P3).

5.2 Accuracy of AdvisingWise Responses

AdvisingWise demonstrated strong accuracy across multiple evaluation methods. Expert validation showed that 84.2% (16 of 19) responses achieved high accuracy scores (4-5), while lower-scoring responses resulted from the system’s conservative approach: it declined to answer when uncertain. Advisors’ perceived accuracy on AI-generated responses varied by question type, with highest confidence for straightforward handbook-explicit queries and more divided perceptions for questions requiring interpretation. In comparative evaluation conducted using LLM-as-a-Judge, our ReAct-style information retrieval system outperformed standard RAG baseline with a 3:1 preference ratio, demonstrating particular

Table 4: Participants in the user study. Two advisors rated the feature as 5 (extremely useful) under the specific conditions. P4 rated it 5, provided that advisors are not involved in the loop and the chatbot is used only by students, with a disclaimer noting that its answers may be inaccurate. P5 rated it 5, considering generative AI highly useful only for handling factual questions.

Participant ID	Advising Role	Prior Use of GenAI Chatbots	Perceived Usefulness of GenAI Chatbots in Advising	Overall Experience of Using AdvisingWise
P1	Staff Advisor	No	1	3
P2	Faculty Advisor	Yes	3	4
P3	Staff Advisor	Yes	3	4
P4	Faculty Advisor	Yes	5	4
P5	Staff Advisor	Yes	5	4
P6	Faculty Advisor	Yes (very little)	2	3
P7	Staff Advisor	Yes	4	4
P8	Faculty Advisor	Yes	3	3

strength on Handbook-Implicit queries where information was not explicitly stated in source documents.

5.2.1 Perceived Accuracy. Some advisors, particularly faculty, generally lacked familiarity with university-wide policies and requirements. Consequently, accuracy in the user study was assessed as perceived accuracy, defined as the extent to which advisors trusted the response and considered it correct.

Figure 6 illustrates that advisors’ perceptions of accuracy varied by question type. For straightforward, fact-based queries (Q1, Handbook-Explicit), all advisors agreed that the AI-generated responses were highly accurate and trustworthy, even among those with no prior experience using generative AI. However, for the remaining questions, their perceptions were more divided.

Advisors evaluated the accuracy of Q2 (Handbook-Implicit) more critically. Q2 posed a more complex query, asking about the appropriate steps to take after failing a particular course. Due to the absence of an explicit answer in the handbook and the need for interpretation based on existing information, the AI-generated response incorporated hedging language such as “may” and explicitly acknowledged uncertainty in interpreting the rules. P1 perceived this uncertainty as “incorrect information” and therefore strongly disagreed (score = 1) with the response’s accuracy. However, not all advisors interpreted uncertain words as misinformation. Although P6 noted that uncertain language diminished the perceived usefulness of the response-stating, “it doesn’t really tell me the answers. ‘Maybe this’, ‘maybe that’ ... the summary is sort of useless,” they still agree with its accuracy (score = 4). Similarly, P5 considered the inclusion of uncertainty acceptable, explaining, “using less sure language, I do that as a human being all the time, so I cannot fault it for that” (P5).

Q3 addressed a topic not covered in the handbook. Most advisors expressed confidence in the accuracy of the AI-generated response and reported trusting it as long as the system explicitly acknowledged its lack of knowledge. However, some advisors had concerns that the inclusion of extraneous information could potentially mislead students, as P4 explained, “if I strictly focus on only the incorrect part, yes, but I do think that students can interpret this incorrectly” (P4).

Q4 was designed to evaluate the AI’s ability to generate personalized responses. Advisors rated its accuracy positively, with most assigning scores of 4–5. P1, however, gave a lower rating (score = 2), citing a preference for more comprehensive output. Because the question concerned degree requirements, P1 felt the response should include all courses within a category rather than a subset limited by response length.

5.2.2 Expert Validation. Of the 20 sampled questions (§4.1), one was excluded from analysis due to the expert’s inability to determine the correct answer, resulting in a final analytical sample of $n = 19$ questions. Of these, 16 responses (84.2%) received high accuracy scores of 4–5, including 12 responses that achieved the maximum score of 5 (extremely accurate). Table 5 presents the distribution of accuracy scores by question type. The remaining 3 responses (15.8%) received low accuracy scores of 1–2, with the following inaccuracies identified:

- (1) One Handbook-Explicit query (score = 1): The system classified the query as off-topic and declined to provide a response.
- (2) One Handbook-Implicit query (score = 1): The system failed to provide information beyond acknowledging the absence of explicit handbook guidance on the topic.
- (3) One Handbook-Unavailable query (score = 2): The system provided misleading information while simultaneously acknowledging uncertainty.

Importantly, low scores in cases (1) and (2) resulted from conservative responses rather than inaccuracies; case (3) offered exploratory information but transparently communicated its limitations.

Table 5: Expert validation results by question type: mean accuracy scores, question count, and standard deviation. Scores based on five-point scale (1=not correct at all, 5=extremely accurate).

Question Type	Mean	Count	Std
Handbook-Explicit	4.42	12	1.16
Handbook-Implicit	4.00	5	1.73
Handbook-Unavailable	3.50	2	2.12
Total	–	19	–

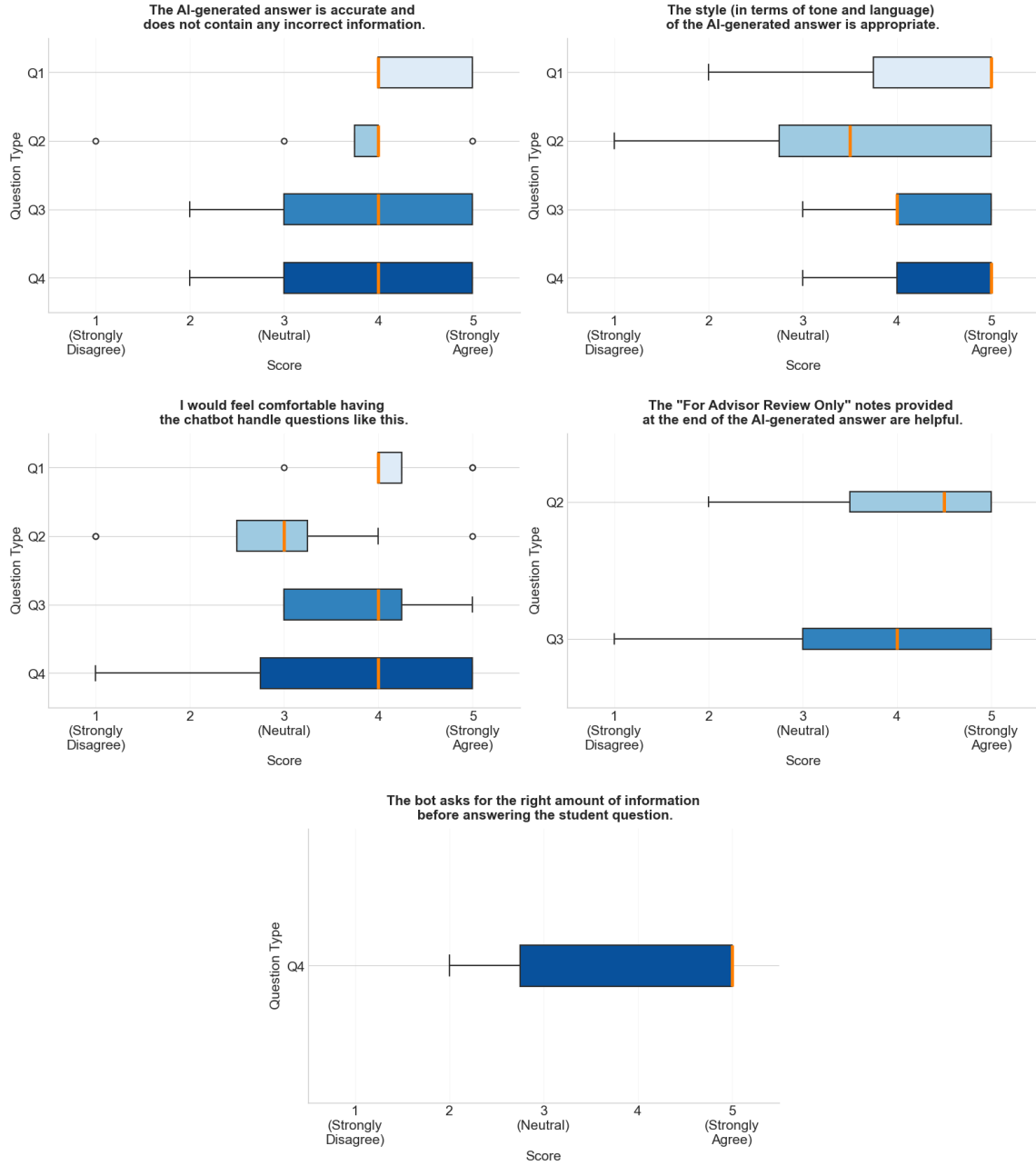


Figure 6: Boxplots illustrating perceived accuracy, response style (in terms of tone and language), and willingness to delegate across different question scenarios

5.2.3 Effectiveness of AdvisingWise’s Information Retrieval System. After excluding 3 queries with expert-validated incorrect responses, the evaluation corpus consisted of 16 queries. To assess position bias and judgment reliability, each query was evaluated twice with system outputs presented in swapped positions. The LLM judge exhibited 75% self-consistency (12/16 queries). Within the self-consistent subset ($n = 12$), preferences were distributed as follows: ReAct-style 50% (6/12), RAG-only 16.7% (2/12), and ties 33.3% (4/12). Excluding

ties, ReAct-style demonstrated a statistically notable 3 : 1 preference ratio over RAG-only approach.

For Handbook-Implicit queries ($n = 3$) in the sample, the ReAct-style system demonstrated superiority, receiving consistent preference in 2 out of 3 cases and achieving a tie in the remaining case. RAG-only was not preferred in any Handbook-Implicit query, suggesting ReAct-style’s iterative retrieval mechanism is better suited for information not explicitly stated in source documents.

Table 6: LLM-as-a-Judge Preference Consistency across Swapped Presentation Orders (n = 16)

Preference Pattern	Count	Pct.
<i>Consistent judgments</i>		
Consistently preferred ReAct-style	6	37.5%
Consistently preferred RAG-only	2	12.5%
Consistently tied	4	25.0%
<i>Subtotal</i>	12	75.0%
<i>Inconsistent judgments</i>		
Switched preference	2	12.5%
Switched to/from tie	2	12.5%
<i>Subtotal</i>	4	25.0%
Total	16	100.0%

5.3 Delegation and Human-Connection

Table 6 shows advisors’ willingness to delegate tasks to AdvisingWise varied significantly based on question complexity and emotional sensitivity, with most willing to delegate straightforward factual queries but reluctant to automate emotionally charged situations. Several advisors, particularly faculty advisors, expressed concerns that increased automation might reduce valuable human connection, which they viewed as central to meaningful advising.

5.3.1 Delegation. Advisors’ willingness to delegate varied across question types. For Q1, a straightforward, fact-based query, 7 of 8 advisors expressed positive attitudes (ratings of 4 or 5) toward delegating factual questions with answers directly available in the handbook. P1, however, rated their willingness as neutral, noting concern that, although the answer was correct in this instance, AdvisingWise might provide incorrect answers in future cases. Given P1’s lack of prior experience with Generative AI and generally cautious view of their reliability, a neutral rating (score = 3) is reasonable.

The majority of participants were neutral or even opposed to delegating Q2 (a Handbook-Implicit query) to the LLM agents. This reluctance was not primarily due to concerns about response quality, but rather the urgency of the situation. Failing a course is a highly critical matter, and advisors preferred more personal interactions with students. As P2 explained:

“If there were five or six nuanced points, you should definitely talk to an advisor at that point, because it already shows how complicated the situation is. It could affect this, it could affect your program completion. At that point, it almost should be ‘click here to contact your advisor’ or something like that” (P2).

For Q3 (Handbook-Unavailable), most advisors (5 of 8) were willing to delegate these questions, valuing the AI’s transparency in acknowledging its limitations. Even P1, generally cautious about LLM agents, supported delegation, implying honesty and acknowledgment of uncertainty increased trust. The remaining 3 advisors, however, raised concerns about the AI providing potentially irrelevant information. While the system attempted to provide additional

details it considered helpful, not all advisors appreciated this approach. P3 remarked, “The details they’re giving do not support the initial answer they gave. They’re separate things.”

While the majority of advisors (5 of 8) expressed positive attitudes toward delegating Q4 (a student-context-dependent question) to AdvisingWise, the remaining advisors raised concerns, noting that the response “relied a little too much on the student [provided information]” (P2). Further discussion is provided in the Personalization section below.

5.3.2 Concerns about Less Human-Connection. While advisors valued AdvisingWise’s efficiency, several expressed concerns about reduced human interaction with students, identifying advising dimensions that extend beyond information delivery. P8 expressed deeper concerns about advising serving as a critical touchpoint:

“It’s hard to meet all of them all the time. Volume is hard. At the same time, this is the only mechanism we have to get our undergrads to talk to us. ... So if that goes away, what’s going to happen? It’s not a good world. That’s my concern, especially for freshmen” (P8).

5.4 AdvisingWise Response Style

Advisors demonstrated diverse preferences for response style, tone, and personalization, reflecting different advising philosophies. Faculty advisors generally preferred concise responses while staff advisors valued comprehensive detail, though these preferences converged toward brevity when AdvisingWise faced uncertainty. Some advisors suggested that emotionally sensitive situations require an adjusted tone, and they also preferred a more proactive approach to gathering information from students.

5.4.1 Style and Tone. Different advising styles lead to varying preferences for the length of AI-generated responses. For example, P4 and P6 consistently complained that the responses were too verbose across almost all question sets, regardless of the type of questions. In contrast, P5 and P7, explicitly appreciated the verbosity, stating,

“The paragraph is how I respond to these, where it’s technically more information than what the student asked for, but the student probably needs that information and doesn’t realize it. So the summary answers the question. There’s stuff in this paragraph that is not related to the question itself, but I still think it is worthwhile that it is included in the response. I really like that” (P5).

“I would probably send all of this, especially because it does answer numerous questions that the student may have. If I was having this conversation with the student, I’d probably start off with academic standing, and then naturally the conversation would go to about degree requirements and next steps. If the student isn’t wanting to have that conversation and wants a quick answer, this sums it up well, so I’d probably choose all of it” (P7).

These preferences appear to be influenced by advisors’ roles. Faculty advisors (P4, P6) generally favored concise responses and expressed

interest in an option to “click to see details”(P6) only when necessary. In contrast, staff advisors (P5, P7) preferred more detailed, verbose responses, emphasizing the value of providing students with comprehensive information.

However, preferences for response length shifted when the AI faced questions it could not fully answer due to handbook limitations. The majority of advisors (P2, P3, P4, P5, P6, P8) preferred a concise response for this kind of questions. Notably, P5, who had favored more detailed responses earlier, opted for brevity in this context, explaining, “unlike in the other ones where it was answering things they didn’t ask and I found it useful, in this case, it feels less useful.”

When addressing Q2, a Handbook-Implicit question, half of the advisors (P2, P3, P5, P8) indicated that the response tone should be adjusted. They emphasized that this recommendation is driven not by the question type itself, but by the specific context in which the question arises. Q2 concerns the possibility of failing a course, a situation with serious implications for the student. Some advisors noted, “The tone is very flat or kind of chill... I would change the tone to create more urgency”(P3). Others emphasized the need for emotional support, with P8 stating:

“The first response would be like to make sure the students okay-what problems did they run into during the semester? Maybe they were sick, maybe someone else was sick, maybe they just feel overwhelmed. That’s the first line response here instead of just providing a bunch of rules as to how to respond the emotions” (P8).

In conclusion, advisors’ preferences regarding the tone and style of AI-generated responses varied, reflecting differences in their individual advising approaches.

5.4.2 Personalization. While advisors appreciated that the system ‘acknowledge[d] that these things (the need for personalization) exist and it asks about additional information’ (P5), feedback indicated that personalization remained too limited. They suggested that other questions, such as Q1 and Q2, could benefit from similar contextualization through sub-questions that gather relevant student information. For example, P3 suggested:

“It can be a case where there’s some user input or a follow-up. The bot can say, ‘These are the rules for how to get co-op credit on your transcript,’ but follow up with questions: ‘What experience do you have? Is this co-op paid? How long have you done this co-op?’ Follow-up questions so the user can say ‘I don’t get paid for this. These are the skills I’ve used for this co-op.’ Then we would be able to tell if this is a ‘go talk to a person, see if there’s a workaround’ situation, or if it’s a general no for everyone” (P3).

Regarding the quality of follow-up questions, most advisors (5 of 8) felt that AdvisingWise collects an appropriate amount of information before generating a response. However, the remaining 3 advisors expressed concern that students might provide inaccurate or incomplete information, which could lead to erroneous responses and resulted in neutral or low ratings (2–3) for the information assessment process. They suggested that accuracy could be improved

by allowing AdvisingWise to access the student information system directly or by requesting that students upload their transcripts.

5.5 UI Design Feedback

Advisors provided mixed feedback on specific interface features, with divided opinions on the value of advisor-only notes and summary placement. While some valued transparency elements like uncertainty warnings, others found them redundant.

5.5.1 Advising-Review-Only Notes. AdvisingWise indicates its limitations by displaying “Advisor Review Only” notes, reminding advisors that certain parts of the AI-generated response may be uncertain. Reactions to this feature were divided. Some advisors appreciated this feature, believing “that’s the right thing to flag too. It’s actually pretty nice”(P2). P5 conveyed strong endorsement, explicitly mentioning:

“This is something that other AIs that I’ve worked with don’t do. They don’t acknowledge when they might be wrong, I love that this could just fix so many problems with ChatGPT-‘I’m not 100% certain, just as a heads up.’ The fact that yours does that is amazing ... If there was a ‘strongly agree times 1 million’ option for the advisory review only, I would choose it” (P5).

However, some advisors viewed it as repetitive. P4 noted, “it’s not like this information is different from what is on the top-this is already there. It would be more useful if it was something that was not said in the answer.” P6 added, “it’s wasting the advisor’s time to even read the advisor section because it says nothing.” P8 further suggested that, if the purpose is to remind advisors, the note would be more effective at the beginning of the response rather than at the end.

5.5.2 Summary. At the end of each response, AdvisingWise consistently provides a brief summary. Similar to feedback on the “Advisor Review Only” notes, advisors expressed diverse perspectives regarding this feature. The majority recommended reversing the order of the summary and the detailed information, placing the summary at the beginning of the response, as “the summary might be the only thing the student reads” (P6). P5, on the other hand, suggested that the placement of the summary should be context-dependent, noting that the optimal order varies across scenarios: “it’s interesting ‘cuz the first one, I think having the summary last made sense, but for that one it made more sense to me for it to be first, so... I don’t know” (P5).

The feedback in this section highlights the importance of balancing advisor control over responses with allowing the LLMs to exercise autonomy. Instead of enforcing a fixed summary format for every response, a more agentic approach may be preferable, enabling the LLMs to determine the order of information or to present the summary only when contextually appropriate.

5.6 Advising Challenges

Although advisors expressed strong enthusiasm for their roles in academic advising, they acknowledged that the current advising workflow presents challenges.

5.6.1 Unfamiliarity of Low-level Details. Faculty advisors are generally less familiar with questions regarding university policies and tend to find these questions less engaging. As P8 noted, “for undergraduates, there are a lot of low-level details about class requirements that I’m not aware of,” and P6 added that figuring out “detailed rules is not the fun part.”

5.6.2 Sheer Volume of Students. Large student caseloads place substantial demands on advisors’ time. P5 noted, “even if the question is super duper easy to answer, like ‘what’s the ad deadline?’ I can answer that in 30 seconds, but if I have 100 people emailing me to ask that, it takes time.” Consequently, advisors spend significant time responding to routine inquiries, which delays students’ access to timely guidance, as P7 explained: “We try to respond as quickly as possible, but when you have a couple hundred students wanting you to reply in 20 minutes, it can get a little complicated.”

5.6.3 Understanding Individual Student Needs. Advisors highlighted the challenge of providing personalized guidance that accounts for each student’s unique academic background, prior experiences, and goals. They emphasized that understanding where a student stands in relation to course content and degree expectations often requires substantial time and contextual knowledge. As P3 explained, “challenges may be gauging where their background lands in concept material ... figuring out their background and sort of customizing their time at [the school name].”

5.6.4 Difficulty Locating Information Efficiently. Even staff advisors, who are typically well-versed in institutional resources, reported difficulties in efficiently locating the necessary information. As P7 noted, “some of the challenges are being able to quickly find the information, see what the requirements are for their specific programs, and then prepare for advising them.” At times, it is also “hard to keep up with the rules” (P6) due to frequent policy changes.

6 Discussion

Based on the results derived from the experiments, we discuss key implications to integrating generative AI tools (AdvisingWise) in academic advising in higher education.

6.1 Design Choices Drive Trust and Adoption

In the study, half of participants (4/8) reported diminished concerns about integrating AI into academic advising following hands-on experience with AdvisingWise. These attitude shifts can be attributed to strategic alignment between system design and anticipated user concerns. During the design phase, we identified key advisor apprehensions and explicitly addressed them through architectural choices. Advisors worried about accuracy: we prioritized it through robust multi-agent response generation architecture and involving advisor review before delivery. Advisors worried about generic responses: we enabled personalization through adaptive student information gathering. The observed positive shifts occurred because advisors could directly observe and evaluate these design decisions in practice, thereby verifying that their concerns had been substantively addressed.

These findings suggest that skepticism toward AI adoption in high-stakes educational domains stems primarily from misalignment between generic AI systems and domain-specific requirements rather than inherent technological limitations. When system design is grounded in domain expert concerns from the outset, direct experience with the system can facilitate the transformation from skepticism to adoption. This underscores the importance of anticipating stakeholder concerns during the design process and validating design-concern alignment through structured pilot programs that enable users to verify whether their needs have been adequately addressed.

6.2 Less Is More When Information Is Unavailable

For questions beyond official documentation (Q3, Handbook-Unavailable), we designed the system to acknowledge uncertainty while providing potentially relevant starting points for advisor research. However, for queries the system could not answer definitively, advisors preferred brief acknowledgment of limitations over lengthy speculative responses, even when those responses included explicit uncertainty disclaimers, as verbose uncertain information increased their review burden. This finding exposes a critical tension in AI system design for high-stakes contexts. When confidence is low, attempting to be helpful through speculative information may undermine trust more than frank admission of knowledge gaps.

6.3 the Need of Accommodating Diverse Advising Styles

Our findings reveal fundamentally different communication philosophies among advisors, challenging the assumption that standardized AI responses can serve all advisors effectively. Preferences diverged across multiple dimensions: some advisors found responses too verbose while others valued comprehensiveness; some preferred directive tones while others favored gentler approaches; some viewed uncertainty warnings as redundant while others valued them as quality assurance. These findings suggest a need for advisor-level customization. Allowing advisors to preset preferences for response length, tone, and detail level could align AI outputs with individual communication styles while reducing manual editing. This points toward a broader insight: effective AI augmentation may require personalization at multiple levels—tailoring not just what information is provided, but how it is communicated to match the human collaborator’s working style.

6.4 Autonomy of LLM and Response Standardization

In the current design, we mandated summaries at the end of every response for consistency and scannability. However, advisors questioned this rigidity: P6 noted summaries sometimes simply repeated earlier content; P5 observed that optimal placement varies by context—sometimes beginning, sometimes end. This feedback exposes a fundamental tension: some decisions cannot be predetermined because they depend on content and context. While advisor preferences (tone, length) can be preset, structural decisions (summary necessity, placement) require situational judgment. This raises a

critical question: how much autonomy should we grant the LLM? Too much risks hallucinations; too little prevents intelligent adaptation. The solution may lie in a tiered autonomy model where the system has high flexibility for presentation choices but low flexibility for factual claims.

6.5 The Confidence Illusion: Why Accuracy Doesn't Predict Delegation

A critical finding challenges conventional assumptions about AI adoption: even when advisors rated AI-generated responses as accurate, they remained reluctant to delegate certain questions. This pattern emerged most clearly in Scenario 2 (student receiving a C+ grade), where advisors assigned high accuracy ratings (predominantly 4/5) yet expressed low comfort with autonomous handling (predominantly 1-3/5). Advisors explained this gap: a failing grade often signals broader student difficulties—health issues, family problems, or personal challenges. While AI accurately communicates policies, it cannot detect underlying issues, provide emotional support, or assess wellbeing. These findings reveal that advising transcends information provision—it requires emotional support and meaningful relationships. This irreplaceable human dimension underscores why human oversight must remain central.

7 Limitations and Future Work

There are several limitations to this study. First, our sample size of user study was relatively small, with only eight advisors from a single department at one institution. All advisors were faculty or staff in the CS department, who may be more familiar with generative AI than advisors in other disciplines. Future work should involve larger, more diverse participant groups across different disciplines and institutional contexts to better validate our findings. Second, our evaluation used hypothetical scenarios with well-structured queries rather than real-world deployment. In practice, students often embed questions within lengthy narratives, provide tangential information, or frame concerns ambiguously. System performance on clear, structured queries may not generalize to these messier, realistic interactions. Third, our study focused solely on advisors' perspectives on AI integration in advising, while students and other important stakeholders (e.g., administrators, institutional leaders) were not included. Future work should include longitudinal deployment studies examining student and advisor interactions over time and larger multi-institutional studies capturing diverse stakeholder perspectives.

8 Conclusion

We introduced AdvisingWise, a multi-agent system that automates information retrieval and response drafting for academic advisors while maintaining human oversight. Our mixed-methods evaluation with expert validation, LLM-as-a-judge benchmarking, and user studies with 8 advisors demonstrated strong effectiveness of our system. Results showed high accuracy and significant improvements over baselines, with advisors valuing transparency, personalization, and workload reduction while preserving human judgment for complex cases. Advisors emphasized that emotional support and mentorship remain irreplaceable despite the system's effectiveness with information-intensive tasks. Future work includes adaptive

formatting to accommodate diverse advising styles, longitudinal deployment, and exploring students' perspectives on AI-advising integration. This work demonstrates human-AI synergy potential in academic advising, and we have only begun to explore what is possible.

9 GenAI Usage Disclosure

Generative AI tools were utilized during the implementation, testing, and evaluation stages of this research. Specifically, generative AI was integrated into the system to support information retrieval, determine next-step actions, and generate responses. Participants interacting with the system were informed of the use of generative AI. In addition, generative AI models were employed to assess the quality of AI-generated responses as part of the evaluation process.

Acknowledgments

We thank Tony Li and all students in the Spring 2025 GenAI for Social Impact class for their valuable feedback and support, and the advisors who participated in the study for their insights. We are grateful to Johnny Redmond for his expert validation of the system responses. This work was partially funded by NSF award 2106797.

References

- [1] Sherif Abdelhamid, James Bangura, Shahryar Shah, et al. 2025. Advisely: AI-Powered Academic Advising Using Large Language Models (LLMs). In *Conference Proceedings, New Perspectives in Science Education 2025*.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).
- [3] Daisuke Akiba and Michelle C Fraboni. 2023. AI-supported academic advising: Exploring ChatGPT's current state and future potential toward student empowerment. *Education Sciences* 13, 9 (2023), 885.
- [4] Anthropic. 2024. Introducing the next generation of Claude. <https://www.anthropic.com/news/claude-3-family>. Accessed: 2025-10-10.
- [5] Maryam Awadh, Mary Jane Best, Muriel Simms, and Bette Lang. 2019. *The Relationship Between the Quality of Academic Advising and the Perception of Academic Outcomes of First-year Saudi College Students*. Ph.D. Dissertation. Advisor(s) Ting-Lan, Ma., AAI22620394.
- [6] Lanthao Benedikt, Chaitanya Joshi, Louisa Nolan, Ruben Henstra-Hill, Luke Shaw, and Sharon Hook. 2020. Human-in-the-loop AI in government: A case study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 488–497.
- [7] Carla E Brodley, Benjamin J Hescott, Jessica Biron, Ali Rensing, Melissa Peiken, Sarah Maravetz, and Alan Mislove. 2022. Broadening participation in computing via ubiquitous combined majors (CS+ X). In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education—Volume 1*. 544–550.
- [8] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW1 (2021), 1–21.
- [9] Tracy Camp, W Richards Adrion, Betsy Bizot, Susan Davidson, Mary Hall, Susanne Hambrusch, Ellen Walker, and Stuart Zweben. 2017. Generation CS: the growth of computer science. *ACM Inroads* 8, 2 (2017), 44–50.
- [10] Heloisa Candello and Michelle Brachman. 2025. Designing conversational agent (ic) systems. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*. 1–4.
- [11] Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung yi Lee. 2024. Large Language Model as an Assignment Evaluator: Insights, Feedback, and Challenges in a 1000+ Student Course. *arXiv:2407.05216 [cs.CL]* <https://arxiv.org/abs/2407.05216>
- [12] C. C. César and M. S. Carvalho. 2011. Stratified sampling design and loss to follow-up in survival models: evaluation of efficiency and bias. *BMC Medical Research Methodology* 11, 99 (2011). doi:10.1186/1471-2288-11-99
- [13] Hai Dang, Karim Benharrah, Florian Lehmann, and Daniel Buschek. 2022. Beyond text generation: Supporting writers with continuous automatic text summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

- [14] Will Epperson, Gagan Bansal, Victor C Dibia, Adam Fournery, Jack Gerrits, Erkan Zhu, and Saleema Amershi. 2025. Interactive debugging and steering of multi-agent ai systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [15] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630 (2024), 625–630. doi:10.1038/s41586-024-07421-0
- [16] Elizabeth Hart-Baldrige. 2020. Faculty Advisor Perspectives of Academic Advising. *NACADA Journal* 40 (07 2020), 10–22. doi:10.12930/NACADA-18-25
- [17] Ye He and Bryant Hutson. 2017. Assessment for Faculty Advising: Beyond the Service Component. *NACADA Journal* 37 (01 2017), 66–75. doi:10.12930/NACADA-16-028
- [18] Owen Henkel, Libby Hills, Adam Boxer, Bill Roberts, and Zach Levonian. 2024. Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability To Mark Short Answer Questions in K-12 Education. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (Atlanta, GA, USA) (L@S '24)*. Association for Computing Machinery, New York, NY, USA, 300–304. doi:10.1145/3657604.3664693
- [19] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. doi:10.1145/3703155
- [20] Omiros Iatrellis, Nicholas Samaras, Konstantinos Kokkinos, and Theodor Panagiotakopoulos. 2024. Leveraging Generative AI for Sustainable Academic Advising: Enhancing Educational Practices through AI-Driven Recommendations. *Sustainability* 16 (09 2024), 7829. doi:10.3390/su16177829
- [21] Johannes Jakubik, Daniel Weber, Patrick Hemmer, Michael Vössing, and Gerhard Satzger. 2023. Improving the efficiency of human-in-the-loop systems: adding artificial to human experts. In *International Conference on Wirtschaftsinformatik*. Springer, 131–147.
- [22] Aaishah Jamaludin, Deepa Alagappan, Wan Fikry, Siti Zainal, Fatin Hadi, Nawal Shaharuddin, and Nurul Rahman. 2021. The Effectiveness of Academic Advising on Student Performance. *International Journal of Advanced Research in Future Ready Learning and Education* 25 (12 2021), 20–29. doi:10.37934/irle.25.1.2029
- [23] Matthew Jeung, Anup Sathya, Wanli Qian, Steven Arellano, Luke Jimenez, and Ken Nakagaki. 2025. Shape n' Swarm: Hands-on, Shape-aware Generative Authoring with Swarm UI and LLMs. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–16.
- [24] Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. Sufficient Context: A New Lens on Retrieval Augmented Generation Systems. arXiv:2411.06037 [cs.CL]. <https://arxiv.org/abs/2411.06037>
- [25] Majeed Kazemitabaar, Justin Chow, Carl Ka To Ma, Barbara J. Ericson, David Weintrop, and Tovi Grossman. 2023. Studying the effect of AI Code Generators on Supporting Novice Learners in Introductory Programming. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 455, 23 pages. doi:10.1145/3544548.3580919
- [26] Mohammad Amin Kuhail, Haseena Alkathreer, Joao Negreiros, Ahmed Seffah, and Omar Alfandi. 2022. Engaging Students With a Chatbot-Based Academic Advising System. *International Journal of Human-Computer Interaction* 39 (06 2022), 1–27. doi:10.1080/10447318.2022.2074645
- [27] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [28] LangChain Inc. [n.d.]. LangGraph. <https://www.langchain.com/langgraph> Accessed: 2025-10-07.
- [29] Kasra Lekan and Zachary A. Pardos. 2024. AI-Augmented Advising: A Comparative Study of GPT-4 and Advisor-based Major Recommendations. In *Proceedings of the 2024 AAAI Conference on Artificial Intelligence (Proceedings of Machine Learning Research, Vol. 257)*. Muktha Ananda, Debshila Basu Malick, Jill Burstein, Lydia T. Liu, Zitao Liu, James Sharpnack, Zichao Wang, and Serena Wang (Eds.). PMLR, 85–96. <https://proceedings.mlr.press/v257/lekan24a.html>
- [30] Kasra Lekan and Zachary A Pardos. 2025. AI-Augmented Advising: A Comparative Study of GPT-4 and Advisor-Based Major Recommendations. *Journal of Learning Analytics* 12, 1 (2025), 110–128.
- [31] 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. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint arXiv:2005.11401* (2021). arXiv:2005.11401 [cs.CL]. <https://arxiv.org/abs/2005.11401>
- [32] Zerui Ma, Michael Hahsler, and Peter Moore. 2025. A Recommender System Architecture for University Curriculum Advising. In *Proceedings of the AAAI Symposium Series, Vol. 5*. 235–241.
- [33] Gonzalo Gabriel Méndez, Luis Galárraga, Katherine Chiluiza, and Patricio Mendoza. 2023. Impressions and strategies of academic advisors when using a grade prediction tool during term planning. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–18.
- [34] Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 2025. Evaluation and benchmarking of llm agents: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 6129–6139.
- [35] Lanlan Mu and Kevin Fossnacht. 2016. Effective advising: How academic advising influences student learning outcomes in different institutional contexts.
- [36] Suchismita Naik, Austin L Toombs, Amanda Snellinger, Scott Saponas, and Amanda K Hall. 2025. Designing with Multi-Agent Generative AI: Insights from Industry Early Adopters. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. 1961–1972.
- [37] Julius Odede and Ingo Frommholz. 2024. JayBot–Aiding University Students and Admission with an LLM-based Chatbot. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*. 391–395.
- [38] Sohyeon Park, Aehong Min, Jesus Armando Beltran, and Gillian R Hayes. 2025. "As an Autistic Person Myself:" The Bias Paradox Around Autism in LLMs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [39] Muhammad Raees, Inge Meijerink, Ioanna Lykourantzou, Vassilis-Javed Khan, and Konstantinos Papangelis. 2024. From explainable to interactive AI: A literature review on current trends in human-AI interaction. *International Journal of Human-Computer Studies* 189 (2024), 103301.
- [40] Rocket.Chat Technologies Corp. 2015. Rocket.Chat: The Communications Platform. <https://github.com/RocketChat/Rocket.Chat>. Accessed: 2025-10-07.
- [41] Logan W Schmidt, Caitlin J Kidder, Ildar Akhmetov, Megan Bebis, Alan C Jamieson, Albert Lionelle, Sarah Maravetz, Sami Rollins, and Ethan Selinger. 2025. An ms in cs for non-cs majors: A ten year retrospective. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*. 1036–1042.
- [42] Krista M. Soria, Elise Kokenge, Cassandra A. Heath, Erin C. Standley, Shannon J.F. Wilson, Jacob R. Connley, and Aaron I. Agramon. 2024. Factors Associated with Academic Advisors' Burnout. *NACADA Journal* 43, 2 (02 2024), 105–120. arXiv:https://meridian.allenpress.com/nacada-journal/article-pdf/43/2/105/3340846/i2330-3840-43-2-105.pdf doi:10.12930/NACADA-23-14
- [43] Karthik Sreedhar, Alice Cai, Jenny Ma, Jeffrey V Nickerson, and Lydia B Chilton. 2025. Simulating cooperative prosocial behavior with multi-agent llms: Evidence and mechanisms for ai agents to inform policy decisions. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 1272–1286.
- [44] Michael Tamascelli, Olivia Bunch, Blake Fowler, Maryam Taeb, and Achraf Cohen. 2025. Academic Advising Chatbot Powered with AI Agent. In *Proceedings of the 2025 ACM Southeast Conference*. 195–202.
- [45] Tavily. 2024. Tavily AI Search API. <https://tavily.com>. Accessed: 2025-10-10.
- [46] Bulut Tok and Gulistan Dogan. 2024. Advisor SeaHawk: An Academic Advisor Chatbot for MSCSIS Students at UNCW. *Journal of Computing Sciences in Colleges* 40, 5 (2024), 138–148.
- [47] Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. 2024. Lave: Llm-powered agent assistance and language augmentation for video editing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 699–714.
- [48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.
- [49] Borui Yang, Md Afif Al Mamun, Jie M. Zhang, and Gias Uddin. 2025. Hallucination Detection in Large Language Models with Metamorphic Relations. *Proc. ACM Softw. Eng.* 2, FSE, Article FSE020 (June 2025), 21 pages. doi:10.1145/3715735
- [50] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*.
- [51] Jingyue Zhang and Ian Arawjo. 2024. ChainBuddy: An AI-assisted Agent System for Helping Users Set up LLM Pipelines. In *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–3.
- [52] Yi Leaf Zhang. 2023. A Dual Perspective on Academic Advising: Challenges Faced by Community College Transfers and Faculty Advisors in Engineering. *NACADA journal* 43, 2 (2023), 72–84.
- [53] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2020, 29 pages.

A Distribution of benchmark questions

Table 7: Distribution of benchmark questions by category and type across 7 frequently asked advising topics. Handbook-Explicit questions have answers explicitly documented in official program handbooks; Handbook-Implicit questions require interpretation or synthesis of handbook content; Handbook-Unavailable questions concern topics, policies, or situations not addressed in official handbooks.

Category Description	Explicit	Implicit	Unavailable	Total
Co-op	10	5	3	18
Core Competency	8	4	3	15
Degree Completion	9	3	3	15
Important Dates	8	3	3	14
Non-CS Courses	5	4	3	12
Online Courses	3	3	3	9
Transfer Credits & Cross-Registration	9	4	4	17
Total	52	26	22	100