JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 12(1), 110–128. https://doi.org/10.18608/jla.2025.8593

AI-Augmented Advising: A Comparative Study of GPT-4 and Advisor-Based Major Recommendations

Kasra Lekan[1]*, Zachary A. Pardos[2]

Abstract

Choosing an undergraduate major is an important decision that impacts academic and career outcomes. In this work, we investigate augmenting personalized human advising for major selection using a large language model (LLM), GPT-4. Through a three-phase survey, we compare GPT suggestions and responses for undeclared first- and second-year students ($n = 33$) to expert responses from university advisors ($n = 25$). Undeclared students were first surveyed on their interests and goals. These responses were then given to both campus advisors and GPT to produce a major recommendation for each student. In the case of GPT, information about the majors offered on campus was added to the prompt. Overall, advisors rated the recommendations of GPT to be highly helpful (4.0 out of 5 on its explanation for the recommendation and 3.8 on its answers to individual student questions) and agreed with its recommendations 33% of the time. Additionally, we observe more agreement with AI's major recommendations when advisors see the AI recommendations before making their own. However, this result was not statistically significant. We categorize qualitative feedback from advisors with an affinity diagram and outline five design implications for future AI-assisted academic advising systems. The results provide a first signal as to the viability of LLMs for personalized major recommendation and shed light on the promise and limitations of AI for advising support.

Notes for Practice

- Large language model (LLM) performance on the major-advising task is likely good enough to aid in the work of human advisors.
- Advisors rated LLM major recommendations and answers to individual student questions favourably.
- Advisors found this intervention to make good use of student-provided information but lacking in its ability to solicit additional follow-up information.

Keywords

Advising, major selection, GPT, LLM, AI-human collaboration, higher education, generative AI, experimental study.

Submitted: 07/08/2024 — **Accepted:** 22/02/2025 — **Published:** 15/03/2025

[1] *Corresponding author Email: kasra.lekan@berkeley.edu Address: School of Education, UC Berkeley, 2121 Berkeley Way, Berkeley, California 94720-1670, USA.

[2] Email: pardos@berkeley.edu Address: School of Education, UC Berkeley, 2121 Berkeley Way, Berkeley, California 94720-1670, USA. ORCID iD: https://orcid.org/0000-0002-6016-7051

1. Introduction

The choice of an undergraduate major is one of the most consequential decisions a student will make in their academic career, affecting earnings (Thomas & Zhang, 2005; Bleemer & Mehta, 2022), job satisfaction (Wolniak & Pascarella, 2005), and degree persistence (Suhre et al., 2007). While some students choose their major independently, many seek advice and recommendations from campus advisors. Academic advising resources vary across institutions, with larger institutions often having substantially greater advisor load (Carlstrom & Miller, 2013).

Recent progress in large language models (LLMs) has drastically increased their ability to comprehend, reason, and generate human language (Ouyang et al., 2022). However, their viability for impactful tasks like assisting with major selection is yet to be explored. Our work aims to fill this gap by evaluating whether LLMs can provide helpful recommendations tailored to individual students' backgrounds and interests regarding their choice of major. This differs from prior natural language processing (NLP) work for student recommendations that focused on automated course planning and scheduling. In our work,

we systematically assessed the strengths and limitations of LLMs to provide personalized guidance on the pivotal decision of which major to pursue. The relationship between demographic factors and major selection is substantiated in higher education research (Wang, 2013; Moakler & Kim, 2014; Wessel et al., 2008). In machine learning, however, demographic factors need to be carefully handled to avoid unintentionally amplifying existing biases (Mehrabi et al., 2022; Bolukbasi et al., 2016).

An undergraduate degree program can be conceptualized as a learning arc that is constrained by degree requirements so that the experience of the degree is more useful than the sum of the individual skills and information learned in each course. Furthermore, major selection is a choice of which learning arc to take, determining the scope of courses, subject matter, and kind of job students will be seen as most qualified for. Academic advising helps determine the optimal learning arc. Under this formulation, a number of research questions arise concerning our study and major advising as a whole.

We investigate the viability of two modern LLMs, GPT-4 and GPT-3.5, to provide major recommendations and question answers at UC Berkeley, a large public university with over 100 majors, by comparing LLM responses to a gold-standard response from professional advisors. Rather than having GPT directly give recommendations to students, our research is designed around the potential for AI to aid advisors in personalizing advice, thus increasing efficiency or student success. We address the following research questions:

> **RQ1:** How closely do the AI's major recommendations, explanations, and question responses match gold-standard advisor responses?
>
> **RQ2:** Does incorporating the student's demographic information affect the AI's performance?
>
> **RQ3:** Does showing AI major recommendations and question answers to advisors influence their own recommendation?

To address these questions, we evaluate AI recommendations against gold-standard advisor responses (RQ1) using both quantitative and qualitative methodologies. We examine the impact of student demographics (RQ2) on LLM recommendations, given the established role of demographics in major selection described in prior literature. Finally, we assess how AI recommendations and answers affect advisor recommendations (RQ3), as this influence is important to the success of AI-assisted academic advising.

The contributions of this work include (1) furthering research on supporting major selection, an important yet understudied area; (2) comparing the relative effectiveness of different LLMs and prompting strategies on the major recommendation task; and (3) determining if LLM-generated recommendations affect subsequent human recommendations.

2. Related Work

Analytics-driven student counselling. The learning analytics field has had a long-standing research thread utilizing analytics, and designing the methodologies to produce them, to guide students in higher education. Course-level analytics have been a focus of past work, with "course signals" at Purdue as an early effort, displaying red, yellow, or green "signals" to students within the learning management system (LMS) based on their current performance (Arnold & Pistilli, 2012). Another early effort from Marist used predictive analytics on course drop-out to notify students of support resources available to them (Jayaprakash et al., 2014). Baucks and colleagues (2024) explored methodologies for producing comparative course analytics, aiming to differentiate between courses by inferred difficulty, and Borchers and Pardos (2023) expanded upon the notion of course credit load to include stress load, time load, and mental effort load, described as course load analytics.

Algorithmic prediction and recommendation for academics. Several authors explore the use of machine learning techniques to predict majors, mostly using clustering techniques (Esteban et al., 2020; Stein et al., 2020; Lang et al., 2022; Maphosa et al., 2024). Much of the work in this area has explored the potential of NLP techniques to provide personalized recommendations and guidance to students navigating their academic trajectories. Jiang and colleagues (2019) presented a method for recommending courses that would best serve as preparation for a specified "goal" course, and Pardos and Jiang (2020) introduced recommendation approaches to broadening students' awareness of lesser-known "serendipitous" courses that may still be of interest to them. Shao and colleagues (2021) introduced PLAN-BERT, a modification of the BERT architecture, to generate personalized multi-semester course plans by incorporating students' past course histories and future courses of interest. Lang and colleagues (2022) extended this approach to major prediction by applying vector embeddings to forecast students' terminal majors based on sequences of courses taken from the beginning of their academic careers. While these studies demonstrate the technical feasibility of AI-powered course planning systems, further research is needed to evaluate their impact on student outcomes and to integrate human advisor expertise with algorithmic recommendations.

The learning analytics field has focused mostly on student-facing analytics in higher education, with advisors not often involved in the intervention. The most related work is that of Ocumpaugh and colleagues (2017), which produced analytics from a web-based tutoring system, ASSISTments, to produce an analytics report for college counsellors. The human-computer interaction (HCI) field has looked more closely at advisor-in-the-loop analytics in higher education, with work exploring the

impact of showing student course grade predictions to advisors during student advising sessions (Méndez et al., 2023; Mendez et al., 2021).

Our work continues to fill this gap of human-algorithmic integration of student and advisor by investigating human advisor expertise combined with AI major recommendations based on student-provided preferences and goals.

Language models in education. Language models, both auto-regressive models like GPT and encoder models like BERT, have been increasingly applied in education settings to personalize assistance to students (Kucirkova et al., 2021; Chang et al., 2022; Pardos & Bhandari, 2023), automate administrative tasks (Bauer et al., 2023; Shaik et al., 2023), or even train teachers (Markel et al., 2023). Many such applications provide positive results but only partially align with the desired outcomes that result when humans perform the task. For instance, Botelho and colleagues (2023) find that encoding student responses for comparison does not capture the breadth of differences that teachers identify when providing feedback to students, and Markel and colleagues (2023) showed that teachers found a benefit from using a simulated student chat system for training but there were limitations in the realism of the scenario. These studies reveal both the promise and limitations of current language models in educational applications, suggesting a need for more sophisticated use of language models to better replicate the nuanced understanding and decision-making processes of human educators.

Human-AI interaction. Research on human-AI collaboration in practical tasks has shown varied outcomes. In their review of multiple studies on AI-supported code translation, Weisz and colleagues (2022) found that out of 10 experiments, only two demonstrated improvements in both efficiency and quality of outcomes when AI was introduced (Desmond et al., 2021; Ashktorab et al., 2021). At the same time, two demonstrated degradation or no change in performance (Weber et al., 2020; F. F. Xu et al., 2022).

The expert-AI collaboration in medicine similarly presents mixed outcomes. Some work shows improvements in accuracy (Tschandl et al., 2020; Reverberi et al., 2022), while other research concludes that experts exhibit confirmation bias or underweight AI predictions (Bashkirova & Krpan, 2024; Agarwal et al., 2023). In an academic advising setting, Méndez and colleagues (2023) investigated the influence of showing predicted grades on the course recommendation strategies of academic advisors. The authors found that advisors rely primarily on their own experience rather than the tool's predictions but spend more time with the tool for lower-performing students.

Effective orchestration of human-AI collaboration remains an open area of research (Capel & Brereton, 2023). Several prior works have examined human-AI interaction, highlighting factors that can impact the effectiveness of the collaboration and user adoption of AI assistance, including transparency, attachment (Gillath et al., 2021), confidence (Chong et al., 2022), and group dynamics (Chiang et al., 2023). The applicability of current findings in the general HCI field to education remains uncertain. Within education, where there is a need for more efficient processes to integrate AI in a humanistic way and humans remain the ultimate decision-makers in most instances, it is essential to gather empirical evidence and examine how AI-human collaboration affects educational tasks.

Our research begins to address the gap in the literature with the experimental application of major recommendations in higher education, drawing from research in human-AI interaction, NLP, and algorithmic prediction in education settings.

3. Methods

3.1 Model Selection

We hypothesized that optimal LLM performance on the task would be determined by the reasoning capabilities of the model and the degree to which responses could be personalized to a student at a specific university rather than an arbitrary university. Thus, we tested GPT-4-0613 (8K token context window) with in-context major names and related department codes and GPT-3.5-Turbo-16K-0613 with in-context major descriptions and related department codes. We also tested fine-tuning a model on university major descriptions and requirements; however, the results were poor since fine-tuning impacted the language generation rather than the reasoning of the model, similar to the findings of Gudibande and colleagues (2023). At the time of conducting this research, GPT-4-32K-0613 (32K token context window) was not available and API (application programming interface) access to the Claude 1 model (approximately 100K token context window) from Anthropic was not widely accessible. Various open-source LLMs, e.g., Llama 2 (Touvron et al., 2023), were candidates for this research. Ultimately, we decided to evaluate GPT models on only the recommendation task and not open-source models to limit the number of research questions we were pursuing with the survey respondents.

We evaluated these model options in terms of their coherence and personalization on five randomly selected student responses. Since the student response dataset is not directly used for analysis in this research but rather is used to facilitate the comparison between AI and advisors, we did not exclude these five randomly selected responses from the student set shown to advisors.

3.2 Major Dataset

The university major dataset was scraped from the university's degree program website. This dataset included descriptions of the majors ($n = 111$) as well as lists of required and elective courses for each major. The dataset was too large to fit in the 16K and 8K context windows available to us. To restrict the size for the 16K context version, we included only the department codes for related courses (e.g., AGRS, ANTHRO, BUDDSTD) and restricted the length of the major description to 600 characters. To shorten the data length for the 8K context version, we included only the department codes for related courses. Despite the limited information we could include in the context, we believed the department codes would enhance model performance since student participants provided department codes for the courses listed as favourites or least favourites.

3.3 Prompt Engineering

Using insights from prompt-engineering research (Zhou et al., 2023; White et al., 2023; Reynolds & McDonell, 2021) and prototyped survey responses written by the authors, we developed a standardized prompt format to ingest each student's survey answers and produce a tailored major recommendation. These prompts were later refined on the five randomly selected student responses used to compare models under consideration.

We wrote four sections for our prompting: (1) system role statement, (2) framing for including major details in the context window, (3) prompts for ingesting the student data, and (4) prompts for retrieving answers to student questions (Figure 1). When writing prompts, we used key best practices from prompt-engineering research (Liu et al., 2023; White et al., 2023). In general, we made the prompts as concise as possible without sacrificing semantic meaning, and we provided clear context for the model's task in the system role, including describing the system's persona (as an "excellent major advisor at [insert_university_name]"), along with the model inputs and outputs. We expressed student data and questions in the form of natural language sentences.

System role statement:

```
You are an excellent major advisor at [insert_university_name]. The following are
 ↪  the majors, along with their descriptions, that you can recommend to students:

<MajorDetails>
# Aerospace Engineering
Related Course Codes: AERO, CIV, COMPSCI, ...

# African American Studies
Related Course Codes: AFRICAM
...
</MajorDetails>
```

Prompt for major recommendation and reasoning:

```
[At least one/Neither] of the student's parents worked in STEM jobs. The student's
 ↪  favorite courses include: [insert courses] The student's least favorite courses
 ↪  include: [insert courses] The student's personal and academic interests include:
 ↪  [insert interests] Potential career paths the student is considering include:
 ↪  [insert career paths]

Based on the student details above, recommend one major. Provide detailed reasoning
 ↪  for why the major is the best fit for the student.
```

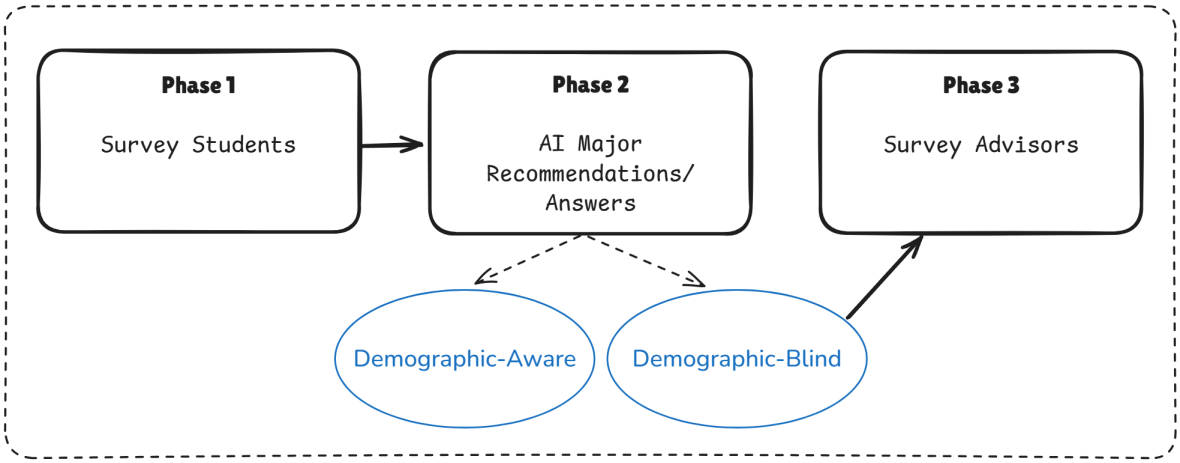
Prompt for student questions:

```
Please answer the following questions from the same student: [insert questions]
```

Figure 1. Finalized prompt formulations. Square brackets represent text to be chosen or replaced using survey responses.

Using the five randomly selected student responses from the survey shown in Figure 3 below, we compared including the major details in the query prompt (separated using XML tags) ("OpenAI Platform", 2023) with including the major details in the system role. We found that including the details in the system role created more personalized recommendations. Next, we compared a third-person prompting strategy, in which the model is provided with details of a student in the third person, and a first-person prompting strategy, in which the user statements are in the first person from a student's perspective as if they are speaking to their advisor. We found that the third-person perspective was more effective as it avoided some GPT safety protocols, thereby improving answer personalization, especially for student questions.

Since we lacked a dataset with which to perform quantitative methods of prompt engineering (Shin et al., 2020), the chosen prompts should be considered a baseline rather than the optimal choices for this task.



	RQ1	RQ2	RQ3	Description
Eval A	✓	✓	✓	Recommendation agreement rate between human advisors and AI
Eval B	✓			Advisor evaluations of AI responses
Eval C	✓	✓		Question answering similarity
Eval D	✓			Recommendation reasoning similarity
Demographics	*Blind*	*Blind/Aware*	*Blind*	"Aware" if the model was prompted with student demographics, "blind" if not

Figure 2. Summary of the methodology (data gathering phases, research questions, and evaluations). For full details, see Section 3, Methods.

3.4 Data Gathering Procedure

We implemented a three-phase data gathering process, summarized in Figure 2, with participants at UC Berkeley. In phase 1 we surveyed students, in phase 2 we generated AI recommendations and answers, and in phase 3 we surveyed advisors. We subsequently describe each phase in further detail. The three-phase procedure was designed to benefit from authentic student responses while minimizing the risks to the students and to test a tool to support advisors rather than supplant them. By surveying students separately in phase 1, we avoid any risks from directly showing them responses. Phases 2 and 3 were designed to help evaluate and compare the AI and advisor responses.

In phase 1, we surveyed a group of undeclared first- and second-year undergraduate students at the university ($n = 33$) using a questionnaire designed to assess factors found to predict success in major programs (e.g., demographics, including gender, ethnicity, and parental STEM occupations) and elicit student details helpful to academic advisors (e.g., coursework preferences, personal interests and strengths, career aspirations). Student responses were solicited from the university classes' Facebook and Reddit pages. The student survey demographic questions (Figure 3) were selected based on insights from prior work on major selection (Wang, 2013; Moakler & Kim, 2014; Wessel et al., 2008), while the background questions were synthesized from questions written by advisors for students in an advising session.

In phase 2, student survey responses were used to generate personalized AI major recommendations and answers to student questions using GPT-4 (13 June 2023 version 0613, 8K token context window), prompted (Figure 1) by including 111 major names to choose from and their related department codes (e.g., ANTHRO, MATH, PSYCH) sourced from their respective major course requirements pages. We also generated recommendations and answers using GPT-3.5 for offline analysis.

In phase 3, students' responses and AI recommendations were provided to university advisors ($n = 25$) as part of a 2×1 between-subjects study design. Each survey form included a single student's data. Advisors were randomly assigned students, and no advisor completed more than two survey forms. Advisors in condition A saw the AI responses after providing their own recommendation, while those in condition B saw the AI response beforehand (Figure 4). This experimental design provides an objective measurement of GPT's effect (Brooks & Hestnes, 2010), which allows us to compare how the AI recommendations influenced advisors, providing insight into human-AI interaction in this context. In the survey, advisors were asked to provide a major recommendation and reasoning as well as answers to the student's questions. The related survey questions contained the

Student survey questions:

1. What is your gender? (based on Wang (2013))
2. What is your ethnicity? Select all that apply. (based on Wang (2013))
3. Did at least one of your parents or guardians have a job in a science, technology, engineering, or math (STEM) field while you were growing up? (based on Moakler and Kim (2014))
4. List 1-2 of your favorite classes that you have taken and why they were your favorite.
5. List 1-2 of your least favorite classes that you have taken and why they were your least favorite.
6. What are your personal interests and academic strengths?
7. What potential career paths are you considering after graduation?
8. What question(s) do you have for an advisor about major selection?

Figure 3. Student survey questions with citations (which were not presented to the students).

Advisor survey questions:

1. [Student background information]
2. *Based on the student details above, recommend one major which is the best fit for the student.*
3. *Provide detailed reasoning for why the major [Selected major] is the best fit for the student.*
4. *Please answer the following questions from the same student: [Student questions]*
5. [AI recommendation and reasoning]
6. Rate the helpfulness of the AI's response to the student. (5-point Likert scale)
7. Please explain your rating of the AI's response.
8. [AI answers to student questions]
9. Rate the helpfulness of the AI's answers to the student's questions. (5-point Likert scale)
10. Please explain your rating of the AI's response.
11. If you have any other feedback or comments about the AI, please include them here.
12. **Based on the student details above, recommend one major which is the best fit for the student.**
13. **Provide detailed reasoning for why that major is the best fit for the student.**
14. **Please answer the following questions from the same student: [Student questions]**

Figure 4. Advisor survey questions. Special formatting corresponds to questions only seen in that condition: **condition A**; **condition B**. Advisors in condition A saw the AI responses after providing their own recommendations, while those in condition B saw the AI response beforehand.

same language used to prompt the LLM (questions 2–4 in condition A, 12–14 in condition B). Additionally, advisors rated the AI's major recommendation, reasoning, and answers. Advisors could optionally provide overall feedback on the AI responses.

3.5 Evaluation

We used four evaluations (A, B, C, and D). The summarized description of these evaluations (Evals) and their relationship with our research questions is shown in Figure 2.

- Eval A is an evaluation of the success of model outputs relative to the advisors' as measured by the rate of agreement between AI and advisor major recommendations. Agreement is the percentage of students for which the model's recommendation matched the advisor's recommendation.
- Eval B is the expert evaluations from advisors on the helpfulness of GPT-4 recommendation and question responses gathered during phase 3 of data gathering.
- Eval C is the cosine similarity between the answers to student questions of the AI and the advisors.
- Eval D is the cosine similarity of the recommendation reasoning in cases where AI and advisor recommendations match.

3.5.1 RQ1: How Closely Do the AI's Major Recommendations, Explanations, and Question Responses Match Gold-Standard Advisor Responses?

RQ1 compares AI performance against expert human performance in the major advising task. Evals A, C, and D, which compare AI responses to advisors', were performed on five models: demographic-blind GPT-4 (the model used for the survey responses), demographic-aware GPT-4, demographic-blind GPT-3.5, demographic-aware GPT-3.5, and a demographic-blind GPT-3.5 restricted to the same 8K context as GPT-4. Demographic-blind models do not have access to the student's gender, ethnicity, or parental STEM occupation.

We compared the similarity of the model outputs to the advisor gold standard using semantic textual similarity measured by cosine similarity between embeddings. The embeddings were generated using all-mpnet-base-v2, a fine-tuned model based on Microsoft's MPNet model (Song et al., 2020), which has performed well on semantic similarity benchmarks (SentenceTransformers, 2024). We hypothesized that the AI recommendations and answers shown to the advisors in condition B (before the advisors wrote their own responses) would be more similar than when advisors wrote their responses independently beforehand. Thus, we used a one-sided T-test to calculate the statistical significance of the embedding differences for each case we are testing.

3.5.2 RQ2: Does Incorporating the Student's Demographic Information Affect the AI's Performance?

We tested whether incorporating the student's ethnicity and gender into the LLM prompt affected the AI's agreement with human advisors in terms of major recommendation and question-answering as measured by Evals A (recommendation agreement) and C (question-answering similarity). We compare these for demographic-blind AI responses, which were shown to the advisors in their survey, and demographic-aware AI responses. Demographic-aware responses were generated by providing the model with the student's gender, ethnicity, or parental STEM occupation in the prompt.

3.5.3 RQ3: Does Showing AI Major Recommendations and Question Answers to Advisors Influence Their Own Recommendation?

RQ3 compares expert human performance against performance by expert humans aided by AI in the major-advising task. We tested the statistical difference in agreement between advisors and the LLMs between conditions A and B (Figure 4). In condition A, the AI response is shown after the advisor provides a recommendation. In condition B, the AI response is shown before the advisor provides a recommendation. The difference in agreement is measured by Eval A (recommendation agreement).

3.5.4 Qualitative Analysis

We explore the advisors' impression of the AI to develop a qualitative evaluation from the advisors of quality (RQ1) and influence (RQ3). In order to quantitatively categorize the advisors' qualitative feedback of the AI's recommendations, we construct an affinity diagram, which categorizes qualitative data for analysis (Lucero, 2015). Responses to questions 7, "Please explain your rating of the AI's response," and 11, "If you have any other feedback or comments about the AI, please include them here," were used (Figure 4) in the affinity diagram to characterize the advisors' feedback about the effectiveness of the AI. Given the relatively small quantity of data and the novelty of the survey questions, a single researcher performed the categorization using categories that emerged from the data. To increase reliability, this researcher first performed a cursory read of the data to construct an initial list of categories. Next, the researcher performed a first pass through the data, adding additional missed categories when needed. After several days, the researcher performed another pass to ensure that all aspects of the responses were represented. Finally, similar categories were grouped together. Note that categories are not mutually exclusive.

4. Results

Among the 33 student participants, 17 were in first year and 16 were in second year. Demographically, 11 were Caucasian, 10 were Asian, eight were Black/African-American, two were Hispanic/Latinx, and two were mixed race. Of the 33 student participants, 21 participants were male, 11 were female, and one identified as "other." All responses were submitted anonymously.

In the phase 3 survey, the 25 advisors were shown responses generated with the GPT-4 demographic-blind model. Offline analysis of that model along with several others demonstrates varying performance on the recommendation, reasoning, and question-answering tasks (Table 1).

4.1 RQ1: How Closely Do the AI's Major Recommendations, Explanations, and Question Responses Match Gold-Standard Advisor Responses?

Overall, advisors viewed the AI's major recommendations, explanations, and question responses favourably. The mean rating for major recommendation and reasoning was 4.0 out of 5, while the mean rating for question-answering and reasoning was 3.8

Table 1. Model performance. Agreement is the percentage of students for which the model's recommendation matched the advisor's recommendation. Major Rec. Reasoning Similarity and Question Response Similarity are the average cosine similarity between the embeddings of the model's and the advisor's responses. The greatest values for each column are bolded.

Model	Agreement Cond. A (AI-2nd)	Agreement Cond. B (AI-1st)	Agreement Overall	Major Rec. Reasoning Similarity	Question Response Similarity
GPT-4 demographic-blind	0.29	**0.38**	**0.33**	0.61	0.51
GPT-4 demographic-aware	**0.41**	0.25	**0.33**	0.61	**0.52**
GPT-3.5 demographic-blind matching 8K context	0.18	0.12	0.15	**0.67**	**0.52**
GPT-3.5 demographic-blind	0.35	0.19	0.27	0.63	0.50
GPT-3.5 demographic-aware	0.35	0.19	0.27	0.65	0.49

out of 5 in terms of helpfulness to students. GPT-4 (demographic-blind) major recommendations to students had an agreement of 33% with the recommendations given by advisors, averaged across both conditions. In many of the disagreement cases, the recommendations from the AI and the advisors were similar, pointing to majors in either the same subject area or the same academic division. Recommendations given by the AI and advisors for the same students are shown in Table 2.

Comparing the similarity of major recommendation reasoning when the AI and advisor agree, GPT-4 demographic-aware had the lowest cosine similarity (0.61), while GPT-3.5 8K demographic-blind had the highest (0.67). Comparing the similarity of answers to student questions, GPT-3.5 demographic-aware had the lowest cosine similarity (0.51), while GPT-3.5 demographic-blind with 8K context had the highest (0.52). Despite having the highest cosine similarity, GPT-3.5 8K demographic-blind was the worst-performing model in terms of recommendation agreement (with an agreement rate of 0.15). The incorporation of major descriptions improved the model's agreement rate by 12%.

4.2 RQ2: Does Incorporating the Student's Demographic Information Affect the AI's Performance?

We observed no differences in overall agreement with the GPT-4 models when student demographics were included versus omitted (Tables 2 and 3). On the question-answering task, the incorporation of background information did not significantly affect the model's semantic similarity with the advisor response (T-stat of 0.24). However, the composition of individual recommendations changed considerably. The GPT-4 demographic-aware model correctly classified two additional students and misclassified two additional students compared to the demographic-blind version, while six other recommendations changed but remained unmatched with the advisor (Table 4). These findings suggest that the integration of demographic information does influence the model, even without a net change in agreement.

4.3 RQ3: Does Showing AI Major Recommendations and Question Answers to Advisors Influence Their Own Recommendation?

To assess whether advisors were influenced by seeing the AI's recommendations, we compared the rate of agreement with the AI's major among advisors in condition A, who were asked to give their responses before being shown the AI's, and in condition B, where they were asked after being shown the AI's answers. We find that there was more agreement in the AI-first condition (0.38) than in the AI-second condition (0.29); however, this difference was not statistically significant ($p = 0.31$).

5. Qualitative Analysis

Given the lack of statistically significant influence of the AI recommendation on the advisors, we further explore the advisors' impression of the AI to develop a qualitative evaluation from the advisors of quality (RQ1) and influence (RQ3). The affinity diagram in Figure 5 summarizes advisor evaluations of the AI's recommendations. Of the 33 total advisor responses, 28 were categorized as "Sufficient/Good," indicating a generally positive reception. Within this category, advisors particularly noted the AI's "Good use of student details" (16 instances). Some advisors (four instances) observed "Similarities to Human Advisor" recommendations, while others commented on the potential for AI to bridge the gap for "Generalist vs. Specialist Advisors" (four instances). However, the diagram also reveals areas for improvement. Under "concerns" (18 instances), advisors most frequently cited the AI's responses as "Reductive/Lacks Nuance" (nine instances). Other concerns included the need for "Asking vs. Telling" (nine instances under "Suggested Improvements"), where advisors emphasized the importance of engaging students in dialogue rather than simply providing recommendations. Additionally, some advisors pointed out that the AI recommended a "High-Frequency Major" (two instances), i.e., one that is more difficult to receive due to its popularity among students, suggesting a need for the system to consider less obvious but potentially suitable options for students.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License (CC BY 4.0)

Table 2. Major recommendations from advisors and LLMs for each student in condition A. Condition A advisors provided their own recommendations first before seeing the AI's. The recommendation from GPT-4 demographic-blind (bolded) was shown to the advisors in the survey.

Condition	Advisor Recommendation	**GPT-4 Demographic-Blind**	GPT-4 Demographic-Aware	GPT-3.5-16K Matching 8K Context Demographic-Blind	GPT-3.5-16K Demographic-Blind	GPT-3.5-16K Demographic-Aware
A	English	Astrophysics	Ancient Greek Roman Studies	Art History	English	Art History
A	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.
A	Applied Mathematics	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.
A	Cognitive Science	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.
A	Nutritional Sciences Toxicology	Psychology	Psychology	Psychology	Psychology	Psychology
A	Environmental Eng. Science	Environmental Science	Environmental Eng. Science	Chemistry	Chemistry	
A	Materials Science Eng. and Business Admin.	Bioengineering	Bioengineering	Bioengineering	Bioengineering	Bioengineering
A	Data Science	Cognitive Science	Data Science	Bioengineering	Data Science	Data Science
A	Mathematics	Applied Mathematics	Applied Mathematics	Applied Mathematics	Applied Mathematics	Applied Mathematics
A	Economics	Applied Mathematics	Comp. Sci.	Applied Mathematics	Applied Mathematics	Applied Mathematics
A	Aerospace Eng.	Aerospace Eng.	Aerospace Eng.	Aerospace Eng.	Aerospace Eng.	Aerospace Eng.
A	Data Science	Data Science	Data Science	Comp. Sci.	Data Science	Data Science
A	Interdisciplinary Studies	English	English	English	English	English
A	Data Science	Data Science	Data Science	Applied Mathematics	Applied Mathematics	Data Science
A	Interdisciplinary Studies	Cognitive Science	Cognitive Science	Cognitive Science	Data Science	Data Science
A	Molecular Cell Biology	Bioengineering	Bioengineering	Bioengineering	Bioengineering	Bioengineering
A	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License (CC BY 4.0)

Table 3. Major recommendations from advisors and LLMs for each student in condition B. Condition B advisors provided their own recommendations after seeing the AI's. The recommendation from GPT-4 demographic-blind (bolded) was shown to the advisor in the survey.

Condition	Advisor Recommendation	**GPT-4 Demographic-Flind**	GPT-4 Demographic-Aware	GPT-3.5-16K matching 8K Context Demographic-Blind	GPT-3.5-16K Demographic-Blind	GPT-3.5-16K Demographic-Aware
B	Integrative Biology	Bioengineering	Bioengineering	Bioengineering	Bioengineering	Molecular Cell Biology
B	Data Science	Applied Mathematics	Comp. Sci.	Applied Mathematics	Applied Mathematics	Comp. Sci.
B	Eng. Math Statistics	Aerospace Eng.	Mechanical Eng.	Aerospace Eng.	Mechanical Eng.	Aerospace Eng.
B	Chemical Biology	Chemical Biology	Chemical Biology	Bioengineering	Bioengineering	Chemistry
B	Legal Studies	Legal Studies	Data Science	Cognitive Science	Economics	Political Economy
B	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.
B	Electrical Eng. Comp. Sci. and Business Admin.	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.
B	Electrical Eng. Comp. Sci. and Business Admin.	Comp. Sci.	Comp. Sci.	Comp. Sci.	French	Comp. Sci.
B	Electrical Eng. Comp. Sci. and Business Admin.	Comp. Sci.	Comp. Sci.	Comp. Sci.	French	Comp. Sci.
B	Political Science	History	African American Studies	History	History	African American Studies
B	Data Science	Cognitive Science	Media Studies	Cognitive Science		Media Studies
B	Data Science	Data Science	Data Science	Applied Mathematics	Data Science	Data Science
B	Chemical Eng. / Materials Science Joint Major	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.
B	Industrial Eng. and Operations Research	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.	Comp. Sci.
B	Astrophysics	Astrophysics	Astrophysics	Astrophysics	Astrophysics	Astrophysics
B	Environmental Economics Policy	Environmental Economics Policy	Economics	Business Admin.	Statistics	Statistics

Table 4. Major recommendations that changed when incorporating demographics into the GPT-4 prompt.

Ethnnicity	Gender	Advisor	GPT-4 demographic-blind	GPT-4 demographic-aware
Caucasian	Female	English	Astrophysics	Ancient Greek Roman Studies
Caucasian	Male	Eng. Math Statistics	Aerospace Eng.	Mechanical Eng.
Caucasian	Male	Data Science	Cognitive Science	Media Studies
Asian	Male	Economics	Applied Mathematics	Comp. Sci.
Latinx	Male	Data Science	Applied Mathematics	Comp. Sci.
African-American	Female	Political Science	History	African American Studies
Latinx	Female	Legal Studies	Legal Studies	Data Science
Asian	Female	Data Science	Cognitive Science	Data Science
Asian	Male	Environ. Economics Policy	Environ. Economics Policy	Economics
African-American	Male	Environ. Eng. Science	Environ. Science	Environ. Eng. Science

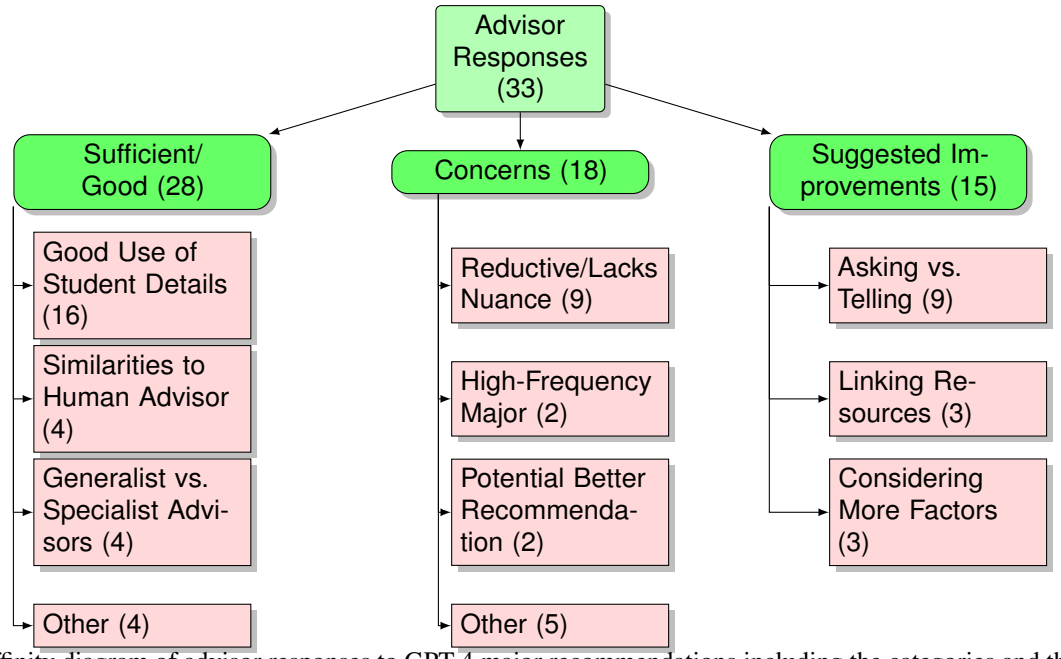


Figure 5. Affinity diagram of advisor responses to GPT-4 major recommendations including the categories and the number of related responses. Categories are not mutually exclusive.

5.1 "Sufficient/Good" Feedback (28 Instances)

Good use of student details (16 instances). Sixteen advisors praised the AI's responses for its thorough consideration of student details and interests. As one commenter noted, "AI pulled a lot of details about the student and was able to apply the details to an explanation for AI's choice." Another highlighted how the AI "provided detailed support for data science as the student's major." The responses were seen as helpful and well reasoned, with one reviewer stating, "I think it took all the student's interests and comments in consideration and gave a good reasoning on the recommendations." Additionally, the AI's ability to address potential concerns was appreciated, as one commenter mentioned: "The last point is helpful around mentioning the concern with math and how that may impact their major."

Similarities to human advisor (four instances). Three advisors noted that the AI's content is similar to what they would say to the student, e.g., "These are all the same things I thought about when reviewing the student's case, as an academic advisor myself." Another went further, noting that the AI "offers more than a simple answer," remarking that the AI answer did not sound forced.

Generalist vs. specialist advisors (four instances). Three advisors discussed how their colleagues specialize in specific schools or degree programs. One advisor observed, "Bioeng[ineering] is a good recommendation here—that one totally slipped my mind! I work with [another school's] students, so that one did not occur to me." This comment illuminates the limitations imposed by an advisor's specialized focus, thereby highlighting the potential value of a well-calibrated AI system in providing a broader range of advising perspectives, at least in an initial interaction with the student. Along these lines, one advisor indicated that the AI would serve only as a generalist advisor: "Once a student has an idea of what major they might be interested in, they would have to go on a dept. website or a human adviser to ask specific questions such as the timeline for declaration. Oftentimes students will ask me about particular faculty, or ease/difficulty of taking certain courses together, which I'm not sure the AI has been taught yet. I recognize that this info is subjective, but students do appreciate the info."

5.2 "Concerns" Feedback (18 instances)

Reductive/lacks nuance (nine instances). Nine advisors critiqued the AI-generated advice for lacking nuance and failing to address the complexities of individual circumstances. As one respondent noted, "The response is great in its reasoning but does not account for the value of an English degree post-graduation," highlighting the AI's inability to consider long-term career implications. Another observed that the AI "doesn't address the student's decision-making process, it's reductive," suggesting a failure to capture the intricate factors influencing educational choices. The limitations of AI in providing personalized guidance were further emphasized: "I think it can be a good exploration tool, but it does not account for the various identities that students have." This sentiment was echoed in the statement that "AI is fine for general information, but lacks nuance and experiential wisdom to provide personalized advice," underscoring the importance of human insight.

High-frequency major (two instances). Two advisors articulated concerns of promoting popular majors at the expense of interdisciplinary or less conventional majors. For instance, one advisor noted, "Computer Science is an impacted major . . . Alternative majors, such as Applied Math, Data Science, Physics, or Cog Sciences might be better alternatives." While this study cannot provide a direct response to such concerns, it should be acknowledged that these dynamics could limit the range of academic paths presented to students. Conversely, one advisor appreciated the system's capacity to suggest less conventional academic routes. The advisor stated, "The recommendation of a lesser-known field may help the student explore other fields and other applications for the degree."

Other (five instances). Other concerns articulated by individual advisors include the efficacy of triaging student emails between an AI and advisors, overly optimistic AI messaging to students, and concerns about the reliability of responses to the same prompt, i.e., "students often ask the same questions over and over, sometimes to the same advisor and sometimes to a variety of trusted advisors, while they deliberate on something like this." One advisor stated that such an advising system should be deployed carefully given its potential for impacting students.

5.3 "Suggested Improvements" Feedback (15 instances)

Asking vs. telling (nine instances). Another recurrent theme that emerged in the feedback around effective advising practices emphasized the necessity of bi-directional dialogue between students and advisors for facilitating informed decision-making. Specifically, one participant underscored the primacy of outlining both advantages and disadvantages: "advising best practice is generally to stick to pros and cons, opportunities and costs [for each potential major]." Additionally, the significance of probing questions was underscored by three advisors. Such questioning can serve to elicit deeper insight into the student's particular decision, as evidenced by the remark, "If it can offer questions to dive [into] the student's interest, that may help solve the student's dilemma on making a decision."

Linking resources (three instances). Three advisors mentioned the potential benefits of incorporating hyperlinks to pertinent resources when delivering recommendations or addressing student queries. As one advisor stated, "A big part of an advisor's job is referring to campus resources. Office hours, student organizations, Career Center, etc." This recommendation aligns with the broader sentiment advocating for an ongoing dialogue between advisors and students, thereby empowering students to make more informed decisions tailored to their individual circumstances.

Considering more factors (three instances). Three advisors referenced various factors (e.g., affiliated school, academic year, student grades, capacity of the major program) that would help constrain the space of possible recommendations. For instance, "Recommending a change of major may not be feasible to a junior student or a transfer student as changing majors that late might delay graduation and degree progress."

6. Discussion

6.1 Key Results and Limitations

Our study revealed both promising potential and important limitations in using LLMs as academic advising assistants. The GPT-4 demographic-blind model achieved moderate agreement (33%) with human advisors' major recommendations, with advisors rating the AI's recommendations and explanations favourably (4.0 of 5 for major recommendations, 3.8 of 5 for question-answering). While incorporating student demographic information altered specific recommendations, it did not significantly impact overall agreement rates with human advisors. However, several methodological limitations warrant consideration: our sample size lacked statistical power for robust between-subjects comparisons, the study was confined to a single institution, and our analysis of AI-advisor agreement may have been inflated by including data from advisors who were exposed to the AI's suggestions before making their own recommendations.

Our finding that advisors were not significantly influenced by AI recommendations is consistent with the findings of Méndez and colleagues (2023) that incorporating predictive analytics into advising sessions did not significantly affect advisors' practices, in part because advisors rely on other important factors besides expected grade. Similarly, we find design implications that include soliciting additional information that may aid in the instructor's decision and the recommendations given to advisors via the AI, as detailed in the section below.

6.2 Design Implications

Our findings suggest several key design implications for AI-assisted academic advising systems:

- **Enhancing nuance and promoting dialogue:** To address concerns about reductive advice, future systems should provide multiple viable options for students to explore, rather than a single recommendation. The AI should ask probing questions to encourage deeper reflection and provide factors for students to consider during their exploration, promoting a more nuanced, dialogue-based interaction. Explainable AI may also play an important role in enhancing nuance. Future research should consider prior work investigating explainable AI, especially in sociotechnical systems as described by Khosravi and colleagues (2022) or on focusing learning analytics dashboards to support equity, potentially through the use of demographic factors as suggested by Williamson and Kizilcec (2022). This design implication stems from advisor feedback in Sections 5.2, Reductive/Lacks Nuance, and 5.3, Asking vs. Telling.

- **Implementing constraints:** To improve recommendation efficacy, the system should incorporate additional factors such as the student's academic year and grades and the capacity of major programs. This could be achieved by modifying the AI's prompt to only list eligible majors based on these constraints; however, a human advisor with direct student experience may be best equipped to handle complex eligibility scenarios. This design implication stems from advisor feedback in Section 5.3, Considering More Factors.

- **Resource integration:** Future iterations should incorporate hyperlinks to relevant resources, addressing the advisors' suggestions for better resource linking. This could include connections to a career engagement office for networking opportunities, professor office hours, student organizations, and department websites. Such integration would provide students with a more comprehensive support system and encourage independent exploration. Future research should consider the designs of previous learning dashboards involving students and advisors such as Millecamp and colleagues (2018), where the authors found that the dashboard should provide most insights at the beginning of a dialogue and that the students desired access to dashboard information even after the interaction ended. This design implication stems from advisor feedback in Section 5.3, Linking Resources.

- **Iterative engagement:** The system should be designed to handle repeated interactions with students, acknowledging that decision-making often involves multiple consultations. This feature would allow students to refine their questions and explore different aspects of their academic choices over time. Future development may draw insight from research on interactivity strategy in intelligent tutoring systems such as in Chi and colleagues (2009). This design implication stems from advisor feedback in Section 5.2, Other.

- **Triage mechanism:** The system could use a triage protocol between AI and human advisors. As suggested by one advisor, the AI could be used when students are more uncertain about their choices, with progression to a human advisor for more specific or complex inquiries. This approach would optimize the use of both AI and human resources in the advising process. Additionally, the triage system could promote student ideation and exploration, potentially in the same spirit of serendipity as the course recommendation system in Pardos and Jiang (2020). This design implication stems from advisor feedback in Section 5.1, Generalist vs. Specialist Advisors.

By incorporating these design implications, we can create an AI-assisted academic advising system that complements human advisors, provides more nuanced and personalized guidance, and better supports students in their major selection process. To evaluate the efficacy of a future design, we propose several viable approaches:

1. maximizing student satisfaction in the medium or long term;
2. minimizing the number of students who fail to declare their major of choice, i.e., minimizing regret; and
3. maximizing advisor efficiency without reducing the two previous metrics.

While this list is not exhaustive, we believe it is a suitable approximation for system-advising efficacy. Future work could investigate these metrics and the trade-offs therein or incorporate learning analytics dashboard evaluation techniques from prior work such as Broos and colleagues (2018), which examined a dashboard for aspiring STEM students and advisors through a questionnaire.

7. Future Work and Limitations

Our study demonstrates the potential for LLMs to serve as intelligent assistants for academic advisors in higher education. However, there are important limitations and ethical considerations that warrant further discussion.

While our study benefited from authentic advisor and student participation, it lacked the statistical power required for a between-subjects two-condition experiment. Future studies should target at least 30 participants per condition to confidently conclude the presence of a significant effect. Additionally, our analysis of the AI's recommendation agreement may have been influenced by including data from condition B advisors, whose own recommendations could have been affected by the model's suggestions. Thus, the measured alignment between human and AI recommendations is potentially inflated.

In practice, some factors would restrict the set of possible major recommendations, e.g., only recommending majors in the College of Letters and Science. We did not limit GPT-4 to recommend majors within a particular division of the university. Taking such restrictions into account would be an interesting step for future work in LLM-based major recommendation. Additionally, while this research focuses on undeclared students at a four-year university, it does not address the needs of prospective transfer students at community colleges whose choices are influenced by their target school.

Our research is confined to the advisors of a single institution. One plausible explanation for the positive orientation of these advisors toward algorithmic collaboration could be their heavy workload, similar to the course credit evaluation staff who were notably receptive to such collaboration as highlighted by L. Xu and colleagues (2023). This is corroborated by one advisor's remark that AI "could be useful seeing as GSAO's [graduate student affairs officers] are overloaded and super busy." Additionally, advisors at this institution may inherently be more open to technological innovations within their field. There is also the possibility of selection bias, as those who chose to participate in this study might have a predisposition toward embracing new technologies. Consequently, future research should investigate whether these findings are consistent across different institutions and varying contexts.

To the extent that advisors consider how well students are cognitively prepared for certain majors (i.e., prerequisite knowledge preparing students for early coursework), LA researchers could investigate the advising input data. Predicting student success is a robust area of research in LA (Jayaprakash et al., 2014; Adnan et al., 2021; Alwarthan et al., 2022) that could be applied to advising. In our research, this was the input data for our prompt and the advisors' survey. Learning assessments to aid both advisors and AI advising could be a ripe area for future research, potentially integrating knowledge tracing or competency evaluation into the AI advising to assess what the student has learned thus far and how that compares to what they are learning in the degree.

In evaluating the LLMs' performance, we opted to use advisor recommendations as the gold standard rather than students' actual major selections. This choice allowed us to test the efficacy of using LLMs to influence advising (RQ3) rather than to influence the student's end major declaration decision. Since major declarations occur at different times in a student's academic journey and interests evolve over time (including through major switches), advisor recommendations offered a more controlled comparison point temporally aligned with the LLM's inputs. This enabled a direct semantic assessment of the LLM's output quality relative to human experts. However, studying the relationship between LLM recommendations, advisor recommendations, and student major selections remains an open direction for future work.

Semantic similarity was a key method used in evaluating the model's responses, which has limitations. First, semantic similarity scores lack interpretability, especially when they are not paired with a clear baseline. Additionally, semantic similarity ultimately relies on the underlying model used to encode the text. Even state-of-the-art models like the one used in this research are insufficient to accurately perform semantic comparison in some instances.

This work also raises the question of where and when AI should fit into the student-advisor relationship. In our study, students provided input to the LLM, which then provided its recommendations and explanations to the advisor, who presumably

might then communicate with the student. Other configurations could be explored, where the LLM conducts a type of triage advising, asking the right questions and follow-up questions of the student before presenting the advisor with recommendations to consider. Alternatively, an advisor could review AI responses and approve some to be sent to the student, while others would need modification or an LLM response would be bypassed in favour of direct communication with the advisor.

In this research, we sought to investigate AI as a tool for helping advisors. Generative AI, even setting aside future advances in the field, has the potential to significantly augment human capabilities in a host of "knowledge work"; however, there is potential for increasing efficiency to cost many jobs (Brynjolfsson et al., 2023; Weidinger et al., 2021). Overall, developing ethical and beneficial applications of LLMs in high-impact domains like education remains an open challenge requiring continued research and awareness of the importance of maintaining human connection in students' educational experiences.

8. Conclusions

This study demonstrates the potential of LLMs like GPT-4 to serve as valuable assistants in academic advising for major selection. By comparing the recommendations generated by GPT-4 with those provided by experienced university advisors, we found that the AI's suggestions were often aligned with expert advice, though not universally so. Advisors and GPT-4 made the exact same major recommendation 33% of the time, with advisors rating GPT-4's justification for its recommendation somewhat highly (4 out of 5 rating of the justification). GPT-4 also appeared up to the task of answering individual student questions related to major selection, with advisors giving a 3.8 out of 5 average rating to those answers. When demographic information was introduced, 10 of the 33 students' major recommendations changed. However, performance as defined by the recommendation agreement between the AI and the advisors was not affected. When analyzing advisors' open-ended feedback on GPT-4's responses, we find that there was consensus ($n = 16$) that the AI made good use of individual student preferences and goal details, which is to say it appears to have excelled in personalization. Advisors also remarked ($n = 9$), however, that, given the students' responses, they would have asked follow-up questions soliciting more information before offering a recommendation. Our findings suggest that while LLMs can significantly augment the advising process, particularly by providing a baseline of recommendations, the integration of human oversight remains crucial. This ensures that the nuanced and personal aspects of advising are maintained, safeguarding against potential biases and errors inherent in automated systems. Due to the largely positive ratings from advisors, the difference in the rate of agreement with the AI in conditions A and B, and qualitative feedback from advisors, LLM recommendations appear to have made a positive impression but did not have a statistically significant influence on advisor recommendations. Overall, the results potentially bode well for human-AI interaction in this area.

Future research should focus on refining the application of these models using the design implications of this work, exploring LLM use across diverse student populations, and investigating further questions about where an AI system should fit into the relationship between student and advisor. Continued collaboration between AI researchers, educational institutions, and advisors will be essential to optimize these tools, ensuring that they enhance the human elements critical to effective academic advising.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work was supported in part by funding from Ascendium Education Group and the Bill & Melinda Gates Foundation. We would like to thank Donna Vivar, Jenae Cohn, and Shawna Dark on behalf of Advising Strategy + Training at UC Berkeley for connecting us to the appropriate campus advisor recruitment channels. This study was approved by the UC Berkeley Committee for the Protection of Human Subjects under IRB Protocol 2023-04-16246.

References

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M., & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access*, *9*, 7519–7539. https://doi.org/10.1109/ACCESS.2021.3049446

Agarwal, N., Moehring, A., Rajpurkar, P., & Salz, T. (2023). Combining human expertise with artificial intelligence: Experimental evidence from radiology [Working Paper, National Bureau of Economic Research]. https://doi.org/10.3386/w31422

Alwarthan, S., Aslam, N., & Khan, I. U. (2022). An explainable model for identifying at-risk student at higher education. *IEEE Access*, *10*, 107649–107668. https://doi.org/10.1109/ACCESS.2022.3211070

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the Second International Conference on Learning Analytics and Knowledge* (LAK 2012), 29 April–2 May 2012, Vancouver, British Columbia, Canada (pp. 267–270). ACM. https://doi.org/10.1145/2330601.2330666

Ashktorab, Z., Desmond, M., Andres, J., Muller, M., Joshi, N. N., Brachman, M., Sharma, A., Brimijoin, K., Pan, Q., Wolf, C. T., Duesterwald, E., Dugan, C., Geyer, W., & Reimer, D. (2021). AI-assisted human labeling: Batching for efficiency without overreliance. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 89:1–89:27. https://doi.org/10.1145/3449163

Bashkirova, A., & Krpan, D. (2024). Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. *Computers in Human Behavior: Artificial Humans*, *2*(1), 100066. https://doi.org/10.1016/j.chbah.2024.100066

Baucks, F., Schmucker, R., Borchers, C., Pardos, Z. A., & Wiskott, L. (2024). Gaining insights into group-level course difficulty via differential course functioning. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (L@S 2024), 18–20 July 2024, Atlanta, Georgia, USA (pp. 165–176). ACM. https://doi.org/10.1145/3657604.3662028

Bauer, E., Greisel, M., Kuznetsov, I., Berndt, M., Kollar, I., Dresel, M., Fischer, M. R., & Fischer, F. (2023). Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda. *British Journal of Educational Technology*, *54*(5), 1222–1245. https://doi.org/10.1111/bjet.13336

Bleemer, Z., & Mehta, A. (2022). Will studying economics make you rich? A regression discontinuity analysis of the returns to college major. *American Economic Journal: Applied Economics*, *14*(2), 1–22. https://doi.org/10.1257/app.20200447

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. D. Lee, U. von Luxburg, R. Garnett, M. Sugiyama, & I. Guyon (Eds.), *Proceedings of the 30th International Conference on Neural Information Processing Systems* (NIPS 2016), 5–10 December 2016, Barcelona, Spain. ACM. https://dl.acm.org/doi/10.5555/3157382.3157584

Borchers, C., & Pardos, Z. A. (2023). Insights into undergraduate pathways using course load analytics. In *Proceedings of the 13th International Conference on Learning Analytics and Knowledge* (LAK 2023), 13–17 March 2023, Arlington, Texas, USA (pp. 219–229). ACM. https://doi.org/10.1145/3576050.3576081

Botelho, A., Baral, S., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2023). Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of Computer Assisted Learning*, *39*(3), 823–840. https://doi.org/10.1111/jcal.12793

Brooks, P., & Hestnes, B. (2010). User measures of quality of experience: Why being objective and quantitative is important. *IEEE Network*, *24*(2), 8–13. https://doi.org/10.1109/MNET.2010.5430138

Broos, T., Verbert, K., Langie, G., Van Soom, C., & De Laet, T. (2018, March). Multi-institutional positioning test feedback dashboard for aspiring students: Lessons learnt from a case study in flanders. In *Proceedings of the Eighth International Conference on Learning Analytics and Knowledge* (LAK 2018), 7–9 March 2017, Sydney, Australia (pp. 51–55). ACM. https://doi.org/10.1145/3170358.3170419

Brynjolfsson, E., Raymond, L., & Li, D. (2023). Generative AI at work [Working Paper 31161, National Bureau of Economic Research]. https://www.nber.org/papers/w31161

Capel, T., & Brereton, M. (2023). What is human-centered about human-centered AI? A map of the research landscape. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, & M. L. Wilson (Eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 23–28 April 2023, Hamburg, Germany (pp. 1–23). ACM. https://doi.org/10.1145/3544548.3580959

Carlstrom, A. H., & Miller, M. A. (2013). 2011 NACADA national survey of academic advising. https://nacada.ksu.edu/Resources/Clearinghouse/View-Articles/2011-NACADA-National-Survey.aspx

Chang, C.-Y., Hwang, G.-J., & Gau, M.-L. (2022). Promoting students' learning achievement and self-efficacy: A mobile chatbot approach for nursing training. *British Journal of Educational Technology*, *53*(1), 171–188. https://doi.org/10.1111/bjet.13158

Chi, M., Jordan, P., Vanlehn, K., & Litman, D. (2009). To elicit or to tell: Does it matter? In *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (AIED 2009), 6–19 July 2009, Brighton, UK (pp. 197–204). IOS Press. https://doi.org/10.3233/978-1-60750-028-5-197

Chiang, C.-W., Lu, Z., Li, Z., & Yin, M. (2023). Are two heads better than one in AI-assisted decision making? Comparing the behavior and performance of groups and individuals in human-AI collaborative recidivism risk assessment. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, & M. L. Wilson (Eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 23–28 April 2023, Hamburg, Germany (pp. 1–18). ACM. https://doi.org/10.1145/3544548.3581015

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, *127*, 107018. https://doi.org/10.1016/j.chb.2021.107018

Desmond, M., Muller, M., Ashktorab, Z., Dugan, C., Duesterwald, E., Brimijoin, K., Finegan-Dollak, C., Brachman, M., Sharma, A., Joshi, N. N., & Pan, Q. (2021). Increasing the speed and accuracy of data labeling through an AI assisted interface. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (IUI 2021), 14–17 April 2021, College Station, Texas, USA (pp. 392–401). ACM. https://doi.org/10.1145/3397481.3450698

Esteban, A., Zafra, A., & Romero, C. (2020). Helping university students to choose elective courses by using a hybrid multi-criteria recommendation system with genetic optimization. *Knowledge-Based Systems*, *194*, 105385. https://doi.org/10.1016/j.knosys.2019.105385

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, *115*, 106607. https://doi.org/10.1016/j.chb.2020.106607

Gudibande, A., Wallace, E., Snell, C., Geng, X., Liu, H., Abbeel, P., Levine, S., & Song, D. (2023). The false promise of imitating proprietary LLMs. *arXiv preprint arXiv:2305.15717*. https://doi.org/10.48550/arXiv.2305.15717

Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, *1*(1), 6–47. https://doi.org/10.18608/jla.2014.11.3

Jiang, W., Pardos, Z. A., & Wei, Q. (2019). Goal-based course recommendation. In *Proceedings of the Ninth International Conference on Learning Analytics and Knowledge* (LAK 2019), 4–8 March 2019, Tempe, Arizona, USA (pp. 36–45). ACM. https://doi.org/10.1145/3303772.3303814

Khosravi, H., Buckingham Shum, S., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074. https://doi.org/10.1016/j.caeai.2022.100074

Kucirkova, N., Gerard, L., & Linn, M. C. (2021). Designing personalised instruction: A research and design framework. *British Journal of Educational Technology*, *52*(5), 1839–1861. https://doi.org/10.1111/bjet.13119

Lang, D., Wang, A., Dalal, N., Paepcke, A., & Stevens, M. L. (2022). Forecasting undergraduate majors: A natural language approach. *AERA Open*, *8*, 233285842211265. https://doi.org/10.1177/23328584221126516

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35. https://doi.org/10.1145/3560815

Lucero, A. (2015). Using affinity diagrams to evaluate interactive prototypes. In J. Abascal, S. Barbosa, M. Fetter, T. Gross, P. Palanque, & M. Winckler (Eds.), *Human-computer interaction–INTERACT 2015. Lecture notes in computer science* (pp. 231–248, Vol. 9297). Springer International Publishing. https://doi.org/10.1007/978-3-319-22668-2_19

Maphosa, M., Doorsamy, W., & Paul, B. (2024). Improving academic advising in engineering education with machine learning using a real-world dataset. *Algorithms*, *17*(2), 85. https://doi.org/10.3390/a17020085

Markel, J. M., Opferman, S. G., Landay, J. A., & Piech, C. (2023). GPTeach: Interactive TA training with GPT-based students. In *Proceedings of the Tenth ACM Conference on Learning @ Scale* (L@S 2023), 20–22 July 2023, Copenhagen, Denmark (pp. 226–236). ACM. https://doi.org/10.1145/3573051.3593393

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6). https://doi.org/10.1145/3457607

Mendez, G., Galárraga, L., & Chiluiza, K. (2021). Showing academic performance predictions during term planning: Effects on students' decisions, behaviors, and preferences. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 8–13 May 2021, Yokohama, Japan. ACM. https://doi.org/10.1145/3411764.3445718

Méndez, G. G., Galárraga, L., Chiluiza, K., & Mendoza, P. (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*, 23–28 April 2023, Hamburg, Germany (pp. 1–18). ACM. https://doi.org/10.1145/3544548.3581575

Millecamp, M., Gutiérrez, F., Charleer, S., Verbert, K., & De Laet, T. (2018). A qualitative evaluation of a learning dashboard to support advisor-student dialogues. In *Proceedings of the Eighth International Conference on Learning Analytics and Knowledge* (LAK 2018), 7–9 March 2018, Sydney, Australia (pp. 56–60). ACM. https://doi.org/10.1145/3170358.3170417

Moakler, M. W., & Kim, M. M. (2014). College major choice in STEM: Revisiting confidence and demographic factors. *The Career Development Quarterly*, *62*(2), 128–142. https://doi.org/10.1002/j.2161-0045.2014.00075.x

Ocumpaugh, J., Baker, R. S., San Pedro, M. O. C. Z., Hawn, M. A., Heffernan, C., Heffernan, N., & Slater, S. A. (2017). Guidance counselor reports of the ASSISTments college prediction model (ACPM). In *Proceedings of the Seventh*

International Conference on Learning Analytics and Knowledge (LAK 2017), 13–17 March 2017, Vancouver, British Columbia, Canada (pp. 479–488). ACM. https://doi.org/10.1145/3027385.3027435

OpenAI Platform. (2023). Retrieved September 21, 2023, from https://platform.openai.com

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*. https://doi.org/10.48550/arXiv.2203.02155

Pardos, Z. A., & Bhandari, S. (2023). Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint arXiv:2302.06871*. https://doi.org/10.48550/arXiv.2302.06871

Pardos, Z. A., & Jiang, W. (2020). Designing for serendipity in a university course recommendation system. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge* (LAK 2020), 23–27 March 2020, Frankfurt, Germany (pp. 350–359). ACM. https://doi.org/10.1145/3375462.3375524

Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., & Cherubini, A. (2022). Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports*, *12*(1), 14952. https://doi.org/10.1038/s41598-022-18751-2

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In Y. Kitamura, A. Quigley, K. Isbister, & T. Igarashi (Eds.), *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 8–13 May 2021, Yokohama, Japan (pp. 1–7). ACM. https://doi.org/10.1145/3411763.3451760

SentenceTransformers. (2024). Pretrained models—Sentence-Transformers documentation. Retrieved February 13, 2024, from https://www.sbert.net/docs/pretrained_models.html

Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2023). Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, *2*, 100003. https://doi.org/10.1016/j.nlp.2022.100003

Shao, E., Guo, S., & Pardos, Z. A. (2021). Degree planning with PLAN-BERT: Multi-semester recommendation using future courses of interest. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), 14920–14929. https://doi.org/10.1609/aaai.v35i17.17751

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020). AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*. https://doi.org/10.48550/arXiv.2010.15980

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*. https://doi.org/10.48550/arXiv.2004.09297

Stein, S. A., M. Weiss, G., Chen, Y., & Leeds, D. D. (2020). A college major recommendation system. In *Proceedings of the 14th ACM Conference on Recommender Systems* (RecSys 2020), 22–26 September 2020, online (pp. 640–644). ACM. https://doi.org/10.1145/3383313.3418488

Suhre, C. J. M., Jansen, E. P. W. A., & Harskamp, E. G. (2007). Impact of degree program satisfaction on the persistence of college students. *Higher Education*, *54*(2), 207–226. https://doi.org/10.1007/s10734-005-2376-5

Thomas, S. L., & Zhang, L. (2005). Post-baccalaureate wage growth within four years of graduation: The effects of college quality and college major. *Research in Higher Education*, *46*(4), 437–459. https://doi.org/10.1007/s11162-005-2969-y

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. https://doi.org/10.48550/arXiv.2302.13971

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H. P., Zalaudek, I., & Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, *26*(8), 1229–1234. https://doi.org/10.1038/s41591-020-0942-0

Wang, X. (2013). Modeling entrance into STEM fields of study among students beginning at community colleges and four-year institutions. *Research in Higher Education*, *54*(6), 664–692. https://doi.org/10.1007/s11162-013-9291-x

Weber, T., Hußmann, H., Han, Z., Matthes, S., & Liu, Y. (2020). Draw with me: Human-in-the-loop for image restoration. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (IUI 2020),17–20 March 2020, Cagliari, Italy (pp. 243–253). ACM. https://doi.org/10.1145/3377325.3377509

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*. https://doi.org/10.48550/arXiv.2112.04359

Weisz, J. D., Muller, M., Ross, S. I., Martinez, F., Houde, S., Agarwal, M., Talamadupula, K., & Richards, J. T. (2022). Better together? An evaluation of AI-supported code translation. In *Proceedings of the 27th International Conference on*

Intelligent User Interfaces (IUI 2022), 22–25 March 2022, Helsinki, Finland (pp. 369–391). ACM. https://doi.org/10.1145/3490099.3511157

Wessel, J. L., Ryan, A. M., & Oswald, F. L. (2008). The relationship between objective and perceived fit with academic major, adaptability, and major-related outcomes. *Journal of Vocational Behavior*, *72*(3), 363–376. https://doi.org/10.1016/j.jvb.2007.11.003

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*. https://doi.org/10.48550/arXiv.2302.11382

Williamson, K., & Kizilcec, R. (2022). A review of learning analytics dashboard research in higher education: Implications for justice, equity, diversity, and inclusion. In *Proceedings of the 12th International Conference on Learning Analytics and Knowledge* (LAK 2022), 21–25 March 2022, online (pp. 260–270). ACM. https://doi.org/10.1145/3506860.3506900

Wolniak, G. C., & Pascarella, E. T. (2005). The effects of college major and job field congruence on job satisfaction. *Journal of Vocational Behavior*, *67*(2), 233–251. https://doi.org/10.1016/j.jvb.2004.08.010

Xu, F. F., Vasilescu, B., & Neubig, G. (2022). In-IDE code generation from natural language: Promise and challenges. *ACM Transactions on Software Engineering and Methodology*, *31*(2), 1–47. https://doi.org/10.1145/3487569

Xu, L., Pardos, Z. A., & Pai, A. (2023). Convincing the expert: Reducing algorithm aversion in administrative higher education decision-making. In *Proceedings of the Tenth ACM Conference on Learning @ Scale* (L@S 2023), 20–22 July 2023, Copenhagen, Denmark (pp. 215–225). ACM. https://doi.org/10.1145/3573051.3593378

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2023). Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*. https://doi.org/10.48550/arXiv.2211.01910