7. Data Appendix

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 1: Student survey questions with citations (that 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 2: Advisor survey questions. Corresponds ***to Version A***; **to Version B**

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.

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-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
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.

Race	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

7.1. Survey Responses

Participant Demographics: Among the 33 student participants, 17 were Freshmen and 16 were Sophomores. Demographically, 11 were Caucasian, 10 were Asian, 8 were Black / African-American, 2 were Hispanic / Latino, and 2 were mixed race. Of the 33 student participants, 21 participants were male, 11 were female, and 1 identified as "Other". All responses were submitted anonymously.

Gathering student responses: Student responses were solicited from Berkeley Class's Facebook and Reddit pages. To maintain the anonymity of student respondents, university email authentication was not required. This, however, resulted in a substantial number of non-useful submissions from our social media recruitment sources. Thus, we filtered out responses that were incoherent (e.g. meaningless form inputs or inputs that did not correspond to the questions), referencing non-university courses, or duplicates of other responses. There were 60, 25, and 24 such responses, respectively, resulting in 78 Phase 1 respondents filtered out.

8. Technical Appendix

8.1. LLM Design Decisions

8.1.1. 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 particular 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, GPT-3.5-Turbo-16K-0613 with in-context major descriptions and related department codes, and GPT-3.5-Turbo-0613 fine-tuned on university major descriptions and requirements. At the time of conducting this research, GPT-4 was not available for fine-tuning and GPT-4-32K-0613 (32K token context window) was not available. Various open-source LLMs, e.g. Llama 2 (Touvron et al., 2023), were candidates for this research. Ultimately, we decided to only evaluate OpenAI GPT models on the recommendation task and not open-source models to limit the number of research questions we were pursuing with the survey respondents.

Among the three candidate models, we had to determine which model to use in the Phase 3 advisor survey while the others would be evaluated through offline analysis. The university major dataset we constructed was scraped from the university's degree program website. This dataset included descriptions of the majors as well as lists of required and elective courses for the major. We noted that for some majors the description did not provide much information unique to the university whereas the related courses were highly specific to the university. For fine-tuning GPT-3.5-Turbo, we were able to use the entire dataset. The dataset, however, 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 only included the department codes for related courses (e.g.ANTHRO, MATH, PSYCH) and restricted the length of the major description to 600 characters. To shorten the data length for the 8K context version, we only included 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 favorites or least favorites.

We evaluated these model options in terms of their coherence and personalization on a set of three 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 three randomly selected responses from the student set shown to advisors.

Fine-tuning: The fine-tuning data consisted of question-answer pairs (Figure 3) intended to associate key details of each major - including required coursework and description - with the corresponding major name. Ultimately, this fine-tuning hindered the LLM, causing it to respond incoherently to student information. Thus, we determined that fine-tuning was not well-suited for our goal of providing the LLM with university-specific major information. Thus, the fine-tuned model was not used to answer our research questions.

Fine-tuning prompt A:

```
User:
    "Describe the major: <major name>"
System:
    "<major description and requirements>"
```

Fine-tuning prompt B:

```
User:
    "<major description and requirements>"
System:
    "<major name>"
```

Figure 3: Prompts for fine-tuning.

8.1.2. Prompt Engineering

Using insights from prompt-engineering research (Zhou et al., 2023; White et al., 2023; Reynolds and McDonell, 2021) and mock survey responses given by the research team, 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 three randomly selected student responses used to compare models under consideration.

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 UC Berkeley") along with the model inputs and outputs. We expressed student data and questions in the form of natural language sentences.

Using the three randomly selected student responses, we compared including the major details in the query prompt (separated using XML tags) (OpenAI) 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 3rd-person prompting strategy, in which the model is provided with details of a student in the 3rd-person, and a 1st-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 3rd-person perspective was more effective as it avoided some unnecessary OpenAI safety language (Ouyang et al., 2022) (Figure 5), thereby improving answer personalization, especially for student questions.

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

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

```
Question: Do I have enough time to complete all my required courses and
graduate by senior spring?

System First-person:
    To determine if you have enough time to complete all your required
    courses and graduate by senior spring,
```
you would need to review the specific requirements of the major(s) you are considering, as well as the course offerings and availability at your university. Additionally, each university may have different credit and graduation requirements. . . . Remember, it's important to regularly communicate with your advisor to ensure you stay on track and to discuss any changes or challenges that may arise. They can provide personalized advice based on your specific situation and available resources at your university.

```
System Third-person:
    Yes, [UniversityName] offers a wide range of courses and resources
    to help you stay on track and complete all your required courses
    within the expected timeframe. ... The number of required courses
    and terms required to complete them may vary depending on your
    chosen major and any additional minors or dual majors you pursue.
    It is important to work closely with your major advisor to create a
    plan that accommodates your interests, allows you to pursue
    internships or research opportunities, and stays on track for
    graduation.\n\nAdditionally, taking advantage of summer courses,
    online courses, or enrolling in a slightly higher course load
    during regular semesters can help you meet your graduation goals.
    [UniversityName] also provides academic support services, such as
    tutoring, study groups, and workshops, to help you succeed
    academically and stay on track. ...
```

Figure 5: First-person versus third-person prompting. The bolded words in the first-person example indicate OpenAI safety language. Sections of the output are omitted using "...".