Background

• The choice of an undergraduate major is one of the most consequential decisions a student will make in their academic career
  • UC Berkeley has nearly 150 majors/minors
• The viability of LLMs for impactful tasks like assisting with major selection is unexplored
• Our work aims to test if LLMs can provide helpful recommendations tailored to individual students’ backgrounds and interests:
  • RQ1: How closely do the AI’s major recommendations, explanations, and question responses match a gold standard advisor response?
  • RQ2: Does incorporating the student’s demographic information affect the AI’s performance?
  • RQ3: Does showing the AI’s response influence an advisor’s subsequent major recommendation?

Prompting Strategy

System role statement:
You are an excellent major advisor at <university name>. The following are the majors, along with their descriptions, that you can recommend to students: ...

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: ...

Chat

Experimental Design

Survey Phase 1
• Surveyed undeclared first and second-year undergraduate students at the university (n=18) eliciting student details helpful to advisors.

Survey Phase 2
• Student survey responses were used to generate personalized AI recommendations for majors and answers to student questions using GPT-4 (June 13th, 2023 version 0613).

Survey Phase 3
• Students' responses and AI recommendations were provided to university advisors (n=18) in 2x1 between-subjects design. Group A saw the AI responses after providing their recommendation, while Group B saw the AI response beforehand.

Preliminary Results

• RQ1: Advisors favorably viewed the AI's major recommendations, explanations, and question responses.
  • Mean rating major rec.: 3.9
  • Mean rating QA: 4.1
• RQ2: marginal differences in agreement in demographic-aware and blind models (0.33 and 0.39)
  • However, half of the students were classified differently between the two scenarios
• RQ3: Substantially more agreement in the AI-1st condition (0.56) than the AI-2nd condition (0.22) – not stat sig.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-4 demographic-blind</td>
<td>0.22</td>
<td>0.56</td>
<td>0.39</td>
<td>0.68</td>
<td>0.53</td>
</tr>
<tr>
<td>GPT-4 demographic-aware</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.67</td>
<td>0.53</td>
</tr>
<tr>
<td>GPT-3.5 demographic-blind matching 8k context</td>
<td>0.11</td>
<td>0.22</td>
<td>0.17</td>
<td>0.77</td>
<td>0.54</td>
</tr>
<tr>
<td>GPT-3.5 demographic-aware</td>
<td>0.33</td>
<td>0.33</td>
<td>0.33</td>
<td>0.67</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Table 1: Model performance. Agreement is the percentage of instances where the model’s recommendation matched the advisor’s. Similarity is the average cosine similarity between explanations.

Condition A Major Recommendations (AI-2nd) Advisor Rec. GPT-4 Rec.
Interdisciplinary Studies Cognitive Science
Applied Mathematics Comp. Sci.
Cognitive Science Comp. Sci.
Mathematics Applied Mathematics
Data Science Cognitive Science
Interdisciplinary Studies English
Molecular Cell Biology BioEng.
Data Science Data Science

Condition B Major Recommendations (AI-1st) Advisor Rec. GPT-4 Rec.
Astrophysics Astrophysics
Data Science Data Science
EE/CS and Business Admin. Comp. Sci.
Environ. Policy Environ. Policy
Legal Studies Legal Studies
Integrative Biology BioEng.
Industrial Eng. and Ops. Comp. Sci.