AI-Augmented Advising: A Comparative Study of ChatGPT-4 and Advisor-based Major Recommendations

Kasra Lekan
University of California, Berkeley
Berkeley, CA 94720
kasra.lekan@berkeley.edu

Zachary A. Pardos
University of California, Berkeley
Berkeley, CA 94720
pardos@berkeley.edu

Abstract

Choosing an undergraduate major is an important decision that impacts academic and career outcomes. We investigate using ChatGPT-4, a state-of-the-art large language model (LLM), to augment human advising for major selection. Through a 3-phase survey, we compare ChatGPT suggestions and responses for undeclared Freshmen and Sophomore students (n=18) to expert responses from university advisors (n=18). Undeclared students were first surveyed on their interests and career goals. These responses were then given to both campus advisors and to ChatGPT to produce a major recommendation for each student. In the case of ChatGPT, information about the majors offered on campus was added to the prompt. Advisors, overall, rated the recommendations of ChatGPT to be highly helpful and agreed with their recommendations 39% of the time. Additionally, we find substantially more agreement with AI major recommendations when advisors see the AI recommendations before making their own. However, this result was not statistically significant, possibly owing to insufficient data collected thus far. 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.

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 and Zhang, 2005, Bleemer and Mehta, 2022], job satisfaction [Wolniak and Pascarella, 2005], and degree persistence [Suhre et al., 2007]. While some students select their major independently, many seek advice from campus advisors for their decision. Academic advising resources vary across institutions with larger institutions often having substantially greater advisor load [Carlstrom and Miller, 2013].

Recent progress in Large Language Models (LLMs) has drastically increased their ability to comprehend, reason with, and generate human language [Ouyang et al., 2022]. However, their viability for impactful tasks like assisting with major selection has been yet unexplored. Our work aims to fill this gap by evaluating if LLMs can provide helpful recommendations tailored to individual students' backgrounds and interests regarding their choice of major. This differs from prior NLP work for student recommendations that focused on automated course planning and scheduling. To our knowledge, no prior work has systematically assessed the strengths and limitations of LLMs for providing personalized guidance on the pivotal decision of which major to pursue.

The premise of this research was to potentially aid advisors in personalizing advice, rather than have ChatGPT directly recommend to students. We investigate the viability of state-of-the-art generative

NeurIPS'23 Workshop on Generative AI for Education (GAIED).

LLMs, ChatGPT-4 and ChatGPT-3.5, to provide major selection assistance at UC Berkeley, a large public university serving 30,000 undergraduates, by comparing LLM responses to a gold-standard response from professional advisors through the following research questions: 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 major recommendation?

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

Recent work has explored the potential of natural language processing (NLP) techniques to provide personalized recommendations and guidance to students navigating their academic trajectories. Shao et al. [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 et al. [2022] extended this approach by applying NLP and vector embeddings specifically to forecast students' terminal majors based on sequences of courses taken from the beginning of their academic careers. Their work demonstrates how even a single initial course selection contains signals predictive of eventual major selection. Méndez et al. [2023] investigated how showing predicted grades influences the course recommendation strategies of academic advisors. In a study using simulated student profiles, they 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. Building on these advances in NLP for academic planning and forecasting, we investigate the potential of ChatGPT to integrate a student's background and interests to offer personalized major recommendations and answer questions.

Language Models in Education:

Language models, both auto-regressive models like GPT and sequence-to-sequence models like BERT, have been increasingly applied in education settings to personalize assistance to students [Kucirkova et al., 2021, Chang et al., 2022, Pardos and Bhandari, 2023], automate administrative tasks [Bauer et al., 2023, Botelho 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 et al. [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 et al. [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. Additionally, notable concerns regarding equity, privacy, and safety arise when using NLP techniques in educational settings [Yan et al., 2023, Sanusi et al., 2023].

Human-AI Interaction:

Effective orchestration of human-AI collaboration remains an open area of research [Capel and 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].

Together, these works showcase different applications of machine learning in education, from automated assessment to course recommendation and teacher training while highlighting the need to carefully design the human-AI interaction.

3 Methods

3.1 Survey Procedure

We implemented a three-phased survey process of participants at UC Berkeley. In Phase 1, we surveyed a group of undeclared first and second-year undergraduate students at the university, with a target n of 35, using a questionnaire designed to assess factors found to predict success in major programs (e.g. demographics and parental STEM occupations) and elicit student details helpful to academic advisors (e.g. coursework preferences, personal interests and strengths, career aspirations). The student survey demographic questions (Figure 2) were selected based on insights from prior work on major selection [Wang, 2013, Moakler and Kim, 2014, Wessel et al., 2008] while the background questions were synthesized from questions written by advisors.

In Phase 2, student survey responses were used to generate personalized AI recommendations for majors and answers to student questions using ChatGPT-4 (June 13th, 2023 version “0613”), prompted (Figure 1) with major names and related department codes (e.g. AGRS, ANTHRO, BUDDSTD). We also generated recommendations and answers using ChatGPT-3.5 for offline analysis. With the larger 16K token context window with ChatGPT-3.5, it was prompted with major names, descriptions, and related department codes.

In Phase 3, students’ responses and AI recommendations were provided to university advisors (n=18) in a 2x1 between-subjects design. Each survey form included a single student’s data. Condition A saw the AI responses after providing their own recommendation, while condition B saw the AI response beforehand (Figure 3). This experimental design provides an objective measurement of ChatGPT’s effect [Brooks and 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 answer the student’s questions. The related survey questions contained the same language used to prompt the LLM. Additionally, advisors rated the AI major recommendation, reasoning, and answers. Advisors could also provide overall feedback on the AI responses.

System role statement:

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

<MajorDetails>
# Major 1
...
</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: ... The student's least favorite courses include:
↪  ... The student's personal and academic interests include: ... Potential
↪  career paths the student is considering include: ...

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: ...
```

Figure 1: Finalized prompt formulations. "..." indicates where text would be inserted (either answers from the student form or major information).

3.2 Evaluation

RQ1: How closely do the AI’s major recommendations, explanations, and question responses match a gold standard advisor response?

During Phase 3, we gathered expert evaluations from advisors (Eval 1) on the helpfulness of the ChatGPT-4 recommendation and question responses. Additionally, we perform offline evaluations of the success of model outputs relative to the advisors based on (Eval 2) the rate of agreement between AI and advisor recommendation, (Eval 3) the similarity of the answers to student questions, and (Eval 4) the similarity of the recommendation reasoning in cases where AI and advisor recommendations match. The offline analyses are performed on demographic-blind and demographic-aware ChatGPT-4 and ChatGPT-3.5 as well as a demographic-blind ChatGPT-3.5 restricted to the same 8k context as ChatGPT-4. All four Evals are used to answer RQ1. With Evals 2, 3, and 4 we report overall results and those restricted to subjects in condition A to control for the influence of the AI's responses on the advisor's major recommendation and reasoning.

We compare 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]. We use a one-sided T-test to calculate the statistical significance of the embedding differences for each case we are testing.

RQ2: Does incorporating the student's demographic information improve the AI's performance?

In Phase 2, we do not prompt the LLM with the student's race and ethnicity (demographic-blind) by default. The relationship between demographic factors and major selection is substantiated in higher education research [Wang, 2013, Moakler and 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]. Investigating the inherent bias in LLMs is a significant and ongoing research area [Feng et al., 2023, Weidinger et al., 2021, Ouyang et al., 2022]. We test if incorporating the student's race and gender into the LLM prompt improves the AI's agreement with human advisors in terms of major recommendation and question answering as measured by Evals 2 and 3.

RQ3: Does showing the AI's response influence an advisor's major recommendation?

We test the statistical difference in agreement between advisors and the LLMs between conditions A and B (Figure 3). 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 2.

4 Results

Out of the target n of 35, we have received 18 responses (Section 7.1 includes demographic details) which capture a preliminary snapshot of our findings. In the survey, advisors were shown responses generated with the ChatGPT-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).

RQ1: How closely do the AI's major recommendations, explanations, and question responses match a gold standard advisor response?

Overall, advisors viewed the AI's major recommendations, explanations, and question responses favorably. The mean rating for the major recommendation and reasoning was 3.9 out of 5 while the mean rating for the question answering and reasoning was 4.1 out of 5 in terms of helpfulness to students. ChatGPT-4 (demographic-blind) major recommendations to students had an agreement of 39% 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, either as majors in 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, ChatGPT-4 demographic-aware had the lowest cosine similarity (0.67) while ChatGPT-3.5 demographic-blind with 8k context had the highest (0.77). Comparing the similarity of answers to student questions, ChatGPT-3.5 demographic-aware had the lowest cosine similarity (0.51) while ChatGPT-3.5 demographic-blind with 8k context had the highest (0.54). Despite having the highest cosine similarity, ChatGPT-3.5 demographic-blind with 8k context was the worst-performing model in terms of recommendation agreement (with an agreement rate of 0.17). The incorporation of major descriptions improved the model's agreement rate by roughly 11%.

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.

Model	Agreement Cond. A (AI-2nd)	Agreement Cond. B (AI-1st)	Agreement Overall	Major Rec. Reasoning Similarity	Question Response Similarity
GPT-4 demographic-blind	0.22	0.56	0.39	0.68	0.53
GPT-4 demographic-aware	0.33	0.33	0.33	0.67	0.53
GPT-3.5 demographic-blind matching 8k context	0.11	0.22	0.17	0.77	0.54
GPT-3.5 demographic-blind	0.22	0.33	0.28	0.69	0.52
GPT-3.5 demographic-aware	0.33	0.33	0.33	0.67	0.51

RQ2: Does incorporating the student's demographic information affect the AI's performance?

We observed marginal differences in agreement with the ChatGPT-4 models when student demographics were included versus omitted, registering agreement rates of 0.33 and 0.39, respectively. 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 -0.016). However, the composition of individual recommendations changed considerably. The ChatGPT-4 demographic-aware model correctly classified four additional students and misclassified five additional students compared to the demographic-blind version. These findings suggest that the integration of demographic information does exert an influence on the model, though the net change in agreement is low.

RQ3: Does showing the AI's response influence an advisor's subsequent major recommendation?

To assess if advisors were influenced by seeing the AI's recommendations, we compared the rate of agreement with the AI's major among advisors in Condition B, who were shown the AI recommendation before being asked to give their own, and in Condition A, where they were asked first. We employed a one-tailed T-test comparing agreement between conditions. We find that there was substantially more agreement in the AI-1st condition (0.56) than in the AI-2nd condition (0.22), however, this difference was just shy of statistical significance (p = 0.08).

5 Discussion

Due to the largely positive ratings from advisors and the difference in the rate of agreement with the AI in conditions A and B, LLM recommendations appear to have made a positive impression and possibly had an influence on advisor recommendations which would bode well for human-AI interaction in this area. This potential is further corroborated by the positive orientation presented in the open-ended feedback from advisors. The source of this positive orientation may be the heavy workload for college advisors, similar to the administrators in Xu et al. [2023] who were more open to algorithmic collaboration due to their heavy workload.

Given the positive impact on agreement that occurred when incorporating the major descriptions into ChatGPT-3.5's prompt, there may be potential for improving AI performance on the advising task with larger context windows. When student demographics were incorporated into ChatGPT-4's prompt, however, half of the major recommendations changed resulting in a slightly lower overall agreement rate. Thus, the employment of LLMs in educational contexts necessitates deliberation over the types of information supplied to the models.

In open-ended feedback left by advisors in the survey, a few articulated that the AI's answers to student questions, especially for broad questions, the AI's responses were more thorough than their own. Other advisors noted, however, that AI answers were somewhat surface-level or lacked nuance such as failing to consider the broader implications of selecting particular majors on job prospects.

Several 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.

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 multiple 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." Advisors also described the practice of posing follow-up questions to the students, illustrating this with, "I would ask [the] student to explain what making it to the top meant to them?"

Another recurrent theme was the potential benefits of incorporating hyperlinks to pertinent resources when delivering recommendations or addressing student queries. 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.

These comments underscore the potential for a more complex specification of the advising problem and the related prompting strategy which could better augment human advising in the future.

6 Limitations

Our study demonstrates the potential for large language models (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 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 major choice is constrained by their intended destination school.

In evaluating the LLM's 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. 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 interesting 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.

Generative AI, even setting aside future advances in the field, has the potential to significantly augment human capabilities in a host of "knowledge work." Several authors express concerns about how this will increase efficiency at the cost of many jobs [Li and Raymond, 2023, Weidinger et al., 2021]. In this research, we sought to investigate AI as a tool for helping advisors. 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 and support in students' educational experiences.

Acknowledgments and Disclosure of Funding

This work was supported in part by funding from Ascendium Education Group and the Bill & Melinda Gates Foundation. This study was approved by the UC Berkeley Committee for the Protection of Human Subjects under IRB Protocol 2023-04-16246.

Table 2: Major recommendations from advisors and LLMs for each student. Condition A saw the AI responses after providing their own recommendation, while condition B saw the AI response beforehand. 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
A	Interdisciplinary Studies	Cognitive Science	Bioengineering	Bioengineering	Bioengineering	Bioengineering
A	Applied Mathematics	Computer Science	Computer Science	Computer Science	Computer Science	Computer Science
A	Cognitive Science	Computer Science	Data Science	Applied Mathematics	Applied Mathematics	Data Science
A	Mathematics	Applied Mathematics	Mechanical Engineering	Aerospace Engineering	Mechanical Engineering	Aerospace Engineering
A	Data Science	Cognitive Science	Data Science	Applied Mathematics	Data Science	Data Science
A	Interdisciplinary Studies	English	Cognitive Science	Cognitive Science	Data Science	Data Science
A	Computer Science	Computer Science	Computer Science	Computer Science	Computer Science	Computer Science
A	Molecular Cell Biology	Bioengineering	Computer Science	Computer Science	Computer Science	Computer Science
A	Data Science	Data Science	Astrophysics	Astrophysics	Astrophysics	Astrophysics
B	Computer Science	Computer Science	English	English	English	English
B	Astrophysics	Astrophysics	Bioengineering	Bioengineering	Bioengineering	Molecular Cell Biology
B	Data Science	Data Science	Economics	Business Administration	Statistics	Statistics
B	Electrical Engineering Computer Sciences and Business Administration	Computer Science	Data Science	Bioengineering	Data Science	Data Science
B	Environmental Economics Policy	Environmental Economics Policy	Computer Science	Computer Science	Computer Science	Computer Science
B	Legal Studies	Legal Studies	Applied Mathematics	Applied Mathematics	Applied Mathematics	Applied Mathematics
B	Engineering Math Statistics	Aerospace Engineering	Computer Science	Computer Science	Computer Science	Computer Science
B	Integrative Biology	Bioengineering	Data Science	Cognitive Science	Economics	Political Economy
B	Industrial Engineering and Operations Research	Computer Science	Computer Science	Computer Science	Computer Science	Computer Science

References

E. Bauer, M. Greisel, I. Kuznetsov, M. Berndt, I. Kollar, M. Dresel, M. R. Fischer, and F. Fischer. 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, 2023. ISSN 1467-8535. doi: 10.1111/bjet.13336. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13336. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13336.

Z. Bleemer and A. Mehta. 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, Apr. 2022. ISSN 1945-7782, 1945-7790. doi: 10.1257/app.20200447. URL https://pubs.aeaweb.org/doi/10.1257/app.20200447.

T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. 2016.

A. Botelho, S. Baral, J. A. Erickson, P. Benachamardi, and N. T. Heffernan. 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, 2023. ISSN 1365-2729. doi: 10.1111/jcal.12793. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jcal.12793. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcal.12793.

P. Brooks and B. Hestnes. User measures of quality of experience: why being objective and quantitative is important. *IEEE Network*, 24(2):8–13, Mar. 2010. ISSN 0890-8044. doi: 10.1109/MNET.2010.5430138. URL http://ieeexplore.ieee.org/document/5430138/.

T. Capel and M. Brereton. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–23, Hamburg Germany, Apr. 2023. ACM. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3580959. URL https://dl.acm.org/doi/10.1145/3544548.3580959.

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

C.-Y. Chang, G.-J. Hwang, and M.-L. Gau. Promoting students' learning achievement and self-efficacy: A mobile chatbot approach for nursing training. *British Journal of Educational Technology*, 53(1):171–188, 2022. ISSN 1467-8535. doi: 10.1111/bjet.13158. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13158. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13158.

C.-W. Chiang, Z. Lu, Z. Li, and M. Yin. 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 *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, Hamburg Germany, Apr. 2023. ACM. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3581015. URL https://dl.acm.org/doi/10.1145/3544548.3581015.

L. Chong, G. Zhang, K. Goucher-Lambert, K. Kotovsky, and J. Cagan. 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, Feb. 2022. ISSN 0747-5632. doi: 10.1016/j.chb.2021.107018. URL https://www.sciencedirect.com/science/article/pii/S0747563221003411.

S. Feng, C. Y. Park, Y. Liu, and Y. Tsvetkov. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. 2023.

O. Gillath, T. Ai, M. S. Branicky, S. Keshmiri, R. B. Davison, and R. Spaulding. Attachment and trust in artificial intelligence. *Computers in Human Behavior*, 115:106607, Feb. 2021. ISSN 0747-5632. doi: 10.1016/j.chb.2020.106607. URL https://www.sciencedirect.com/science/article/pii/S074756322030354X.

N. Kucirkova, L. Gerard, and M. C. Linn. Designing personalised instruction: A research and design framework. *British Journal of Educational Technology*, 52(5):1839–1861, 2021. ISSN 1467-8535. doi: 10.1111/bjet.13119. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13119. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13119.

D. Lang, A. Wang, N. Dalal, A. Paepcke, and M. L. Stevens. Forecasting Undergraduate Majors: A Natural Language Approach. *AERA Open*, 8:233285842211265, Jan. 2022. ISSN 2332-8584, 2332-8584. doi: 10.1177/23328584221126516. URL http://journals.sagepub.com/doi/10.1177/23328584221126516.

D. Li and L. Raymond. Erik Brynjolfsson Stanford & NBER. 2023.

J. M. Markel, S. G. Opferman, J. A. Landay, and C. Piech. GPTeach: Interactive TA Training with GPT-based Students. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, pages 226–236, Copenhagen Denmark, July 2023. ACM. ISBN 9798400700255. doi: 10.1145/3573051.3593393. URL https://dl.acm.org/doi/10.1145/3573051.3593393.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–35, July 2022. ISSN 0360-0300, 1557-7341. doi: 10.1145/3457607. URL https://dl.acm.org/doi/10.1145/3457607.

M. W. Moakler and M. M. Kim. College Major Choice in STEM: Revisiting Confidence and Demographic Factors. *The Career Development Quarterly*, 62(2):128–142, June 2014. ISSN 08894019. doi: 10.1002/j.2161-0045.2014.00075.x. URL https://onlinelibrary.wiley.com/doi/10.1002/j.2161-0045.2014.00075.x.

G. G. Méndez, L. Galárraga, K. Chiluiza, and P. Mendoza. 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*, pages 1–18, Hamburg Germany, Apr. 2023. ACM. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3581575. URL https://dl.acm.org/doi/10.1145/3544548.3581575.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, Mar. 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

Z. A. Pardos and S. Bhandari. Learning gain differences between ChatGPT and human tutor generated algebra hints, Feb. 2023. URL http://arxiv.org/abs/2302.06871. arXiv:2302.06871 [cs].

I. T. Sanusi, S. S. Oyelere, H. Vartiainen, J. Suhonen, and M. Tukiainen. A systematic review of teaching and learning machine learning in K-12 education. *Education and Information Technologies*, 28(5):5967–5997, May 2023. ISSN 1360-2357, 1573-7608. doi: 10.1007/s10639-022-11416-7. URL https://link.springer.com/10.1007/s10639-022-11416-7.

T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan. Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, 2:100003, Mar. 2023. ISSN 2949-7191. doi: 10.1016/j.nlp.2022.100003. URL https://www.sciencedirect.com/science/article/pii/S2949719122000036.

E. Shao, S. Guo, and Z. A. Pardos. 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, May 2021. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v35i17.17751. URL https://ojs.aaai.org/index.php/AAAI/article/view/17751.

K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu. MPNet: Masked and Permuted Pre-training for Language Understanding, Nov. 2020. URL http://arxiv.org/abs/2004.09297. arXiv:2004.09297 [cs].

C. J. M. Suhre, E. P. W. A. Jansen, and E. G. Harskamp. Impact of degree program satisfaction on the persistence of college students. *Higher Education*, 54(2):207–226, June 2007. ISSN 0018-1560, 1573-174X. doi: 10.1007/s10734-005-2376-5. URL http://link.springer.com/10.1007/s10734-005-2376-5.

S. L. Thomas and L. Zhang. 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, June 2005. ISSN 0361-0365, 1573-188X. doi: 10.1007/s11162-005-2969-y. URL `http://link.springer.com/10.1007/s11162-005-2969-y`.

X. Wang. 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, Sept. 2013. ISSN 0361-0365, 1573-188X. doi: 10.1007/s11162-013-9291-x. URL `http://link.springer.com/10.1007/s11162-013-9291-x`.

L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, and I. Gabriel. Ethical and social risks of harm from Language Models, Dec. 2021. URL `http://arxiv.org/abs/2112.04359`. arXiv:2112.04359 [cs].

J. L. Wessel, A. M. Ryan, and F. L. Oswald. The relationship between objective and perceived fit with academic major, adaptability, and major-related outcomes. *Journal of Vocational Behavior*, 72(3):363–376, June 2008. ISSN 00018791. doi: 10.1016/j.jvb.2007.11.003. URL `https://linkinghub.elsevier.com/retrieve/pii/S0001879107001005`.

G. C. Wolniak and E. T. Pascarella. The effects of college major and job field congruence on job satisfaction. *Journal of Vocational Behavior*, 67(2):233–251, Oct. 2005. ISSN 00018791. doi: 10.1016/j.jvb.2004.08.010. URL `https://linkinghub.elsevier.com/retrieve/pii/S0001879105000199`.

L. Xu, Z. A. Pardos, and A. Pai. Convincing the Expert: Reducing Algorithm Aversion in Administrative Higher Education Decision-making. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, pages 215–225, Copenhagen Denmark, July 2023. ACM. ISBN 9798400700255. doi: 10.1145/3573051.3593378. URL `https://dl.acm.org/doi/10.1145/3573051.3593378`.

L. Yan, L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, and D. Gašević. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, n/a(n/a), 2023. ISSN 1467-8535. doi: 10.1111/bjet.13370. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13370`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13370.

7 Appendix

7.1 Survey

Participant Demographics: Among the 18 student participants, 9 were Freshmen and 9 were Sophomores. Of the 18 student participants, 6 were Caucasian, 6 were Asian, 3 were Black / African-American, 2 were mixed, and 1 was Hispanic / Latino. Of the 18 student participants, 13 participants were male, 4 were female, and 1 identified as "Other". 18 university advisors participated in the research. All responses from advisors and students were submitted anonymously.

Filtering student responses: In order to maintain anonymity for student respondents, university email authentication was not collected. 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 previous responses. There were 60, 25, and 24 of these respectively.

7.2 Figures

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 (course ID and title) that you have taken and why they were your favorite.
5. List 1-2 of your least favorite classes (course ID and title) 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 2: 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 3: Advisor survey questions. ***Corresponds to Version A***. **Corresponds to Version B**