Proceedings of Machine Learning Research 38:1–12, 2024 AI4ED Workshop

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

Kasra Lekan KASRA.LEKAN@BERKELEY.EDU
and **Zachary A. Pardos** PARDOS@BERKELEY.EDU
University of California, Berkeley

Abstract

Choosing an undergraduate major is an important decision that impacts academic and career outcomes. We investigate using GPT-4, a state-of-the-art large language model (LLM), to augment human advising for major selection. Through a 3-phase survey, we compare GPT suggestions and responses for undeclared Freshmen and Sophomore 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 to 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. Advisors, overall, rated the recommendations of GPT to be highly helpful and agreed with their recommendations 33% of the time. Additionally, we observe more agreement with AI major recommendations when advisors see the AI recommendations before making their own. However, this result was not statistically significant. 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.

Keywords: Advising, major selection, GPT, LLM, AI-Human collaboration, higher education, Generative AI, Experimental study

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 natural language processing (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 GPT directly recommend to students. We investigate the viability of state-of-the-art generative LLMs, GPT-4 and GPT-3.5, to provide major selection assistance at UC Berkeley, a large public university with over 100 majors, by comparing LLM responses to a gold-standard response from professional advisors through pursuing 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 responses?

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 vector embeddings to forecast students' terminal majors based on sequences of courses taken from the beginning of their academic careers. 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.

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 and 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 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.

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 (n=33) 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 (Appendix Figure 1) 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 major recommendations and answers to student questions using GPT-4 (June 13th, 2023 version "0613", 8K token context window), prompted (Appendix Figure 4) to include 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. Given the larger 16K token context available at the time with GPT-3.5, the model was prompted with major names, descriptions, and related department codes.

In Phase 3, students' responses and AI recommendations were provided to university advisors (n=25) as part of a 2x1 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 condition B saw the AI response beforehand (Appendix Figure 2). This experimental design provides an objective measurement of GPT'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 answers to 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.

3.2. Evaluation

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

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

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 highly on semantic similarity benchmarks (SentenceTransformers). We used a one-sided T-test to calculate the statistical significance of the embedding differences for each case we are testing.

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

In Phase 2, we did 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 tested if incorporating the student's race 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 2 and 3.

RQ3: Does showing AI major recommendations and question answers to advisors influence their own responses?

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

We collected responses from 33 students (Section 7.1 includes demographic details). 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).[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 favorably. The mean rating for the major recommendation and reasoning was 4.0 out of 5 while the mean rating for the question answering and reasoning was 3.8 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, 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 Appendix Table 3.

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%.

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 (Appendix Table 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 (Appendix Table 4). These findings suggest that the integration of demographic information does exert an influence on the model, even without a net change in agreement.

1. A work-in-progress report was presented at a non-archival workshop at the midpoint of data collection (n=18).

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	0.29	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

RQ3: Does showing AI major recommendations and question answers to advisors influence their own responses?

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 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-1st condition (0.38) than in the AI-2nd condition (0.29), however, this difference was not statistically significant ($p = 0.31$).

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 bodes 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 of advisors, similar to that of course credit evaluation staff who were similarly open to algorithmic collaboration (Xu et al., 2023).

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

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]." 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.

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 UC Berkeley. 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.

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. 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.

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 increasing 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 in students' educational experiences.

Acknowledgments

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

Elisabeth Bauer, Martin Greisel, Ilia Kuznetsov, Markus Berndt, Ingo Kollar, Markus Dresel, Martin R. Fischer, and Frank 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.

Zachary Bleemer and Aashish 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, April 2022. ISSN 1945-7782, 1945-7790. doi: 10.1257/app.20200447. URL https://pubs.aeaweb.org/doi/10.1257/app.20200447.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. 2016.

Anthony Botelho, Sami Baral, John A. Erickson, Priyanka Benachamardi, and Neil 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.

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

Tara Capel and Margot 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, April 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.

Ching-Yi Chang, Gwo-Jen Hwang, and Meei-Ling 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.

Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming 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, April 2023. ACM. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3581015. URL https://dl.acm.org/doi/10.1145/3544548.3581015.

Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan 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, February 2022. ISSN 0747-5632. doi: 10.1016/j.chb.2021.107018. URL https://www.sciencedirect.com/science/article/pii/S0747563221003411.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. 2023.

Omri Gillath, Ting Ai, Michael S. Branicky, Shawn Keshmiri, Robert B. Davison, and Ryan Spaulding. Attachment and trust in artificial intelligence. *Computers in Human Behavior*, 115:106607, February 2021. ISSN 0747-5632. doi: 10.1016/j.chb.2020.106607. URL https://www.sciencedirect.com/science/article/pii/S074756322030354X.

Natalia Kucirkova, Libby Gerard, and Marcia 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.

David Lang, Alex Wang, Nathan Dalal, Andreas Paepcke, and Mitchell L. Stevens. Forecasting Undergraduate Majors: A Natural Language Approach. *AERA Open*, 8:233285842211265, January 2022. ISSN 2332-8584, 2332-8584. doi: 10.1177/23328584221126516. URL http://journals.sagepub.com/doi/10.1177/23328584221126516.

Danielle Li and Lindsey Raymond. Erik Brynjolfsson Stanford & NBER. 2023.

Julia M. Markel, Steven G. Opferman, James A. Landay, and Chris 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.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram 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.

Martin W. Moakler and Mikyong Minsun 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.

Gonzalo Gabriel Méndez, Luis Galárraga, Katherine Chiluiza, and Patricio 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, April 2023. ACM. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3581575. URL https://dl.acm.org/doi/10.1145/3544548.3581575.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

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

SentenceTransformers. Pretrained Models — Sentence-Transformers documentation. URL https://www.sbert.net/docs/pretrained_models.html.

Thanveer Shaik, Xiaohui Tao, Christopher Dann, Haoran Xie, Yan Li, and Linda Galligan. Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, 2:100003, March 2023. ISSN 2949-7191. doi: 10.1016/j.nlp.2022.100003. URL https://www.sciencedirect.com/science/article/pii/S2949719122000036.

Erzhuo Shao, Shiyuan Guo, and Zachary 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.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and Permuted Pre-training for Language Understanding, November 2020. URL http://arxiv.org/abs/2004.09297. arXiv:2004.09297 [cs].

Cor J. M. Suhre, Ellen P. W. A. Jansen, and Egbert 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.

Scott L. Thomas and Liang 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.

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

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from Language Models, December 2021. URL http://arxiv.org/abs/2112.04359. arXiv:2112.04359 [cs].

Jennifer L. Wessel, Ann Marie Ryan, and Frederick 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.

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

Lingrui Xu, Zachary A. Pardos, and Anirudh 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.

7. Appendix

A complete appendix can be found in our repository.

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.

7.2. Prompting

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