Modulating Language Model Experiences through Frictions

Katherine M. Collins[1], Valerie Chen[2], Ilia Sucholutsky[3], Hannah Rose Kirk[4], Malak Sadek[5], Holli Sargeant[1], Ameet Talwalkar[2], Adrian Weller [1,6], Umang Bhatt[6,7]

[1]University of Cambridge, [2]Carnegie Mellon University, [3]Princeton University, [4]University of Oxford, [5]Imperial College, [6]The Alan Turing Institute, [7]New York University
Correspondence: kmc61@cam.ac.uk umangbhatt@nyu.edu

Abstract

Language models are transforming the ways that their users engage with the world. Despite impressive capabilities, over-consumption of language model outputs risks propagating unchecked errors in the short-term and damaging human capabilities for critical thinking in the long-term. How can we develop scaffolding around language models to curate more appropriate use? We propose *selective frictions* for language model experiences, inspired by behavioral science interventions, to dampen misuse. Frictions involve small modifications to a user's experience, e.g., the addition of a button impeding model access and reminding a user of their expertise relative to the model. Through a user study with real humans, we observe shifts in user behavior from the imposition of a friction over LLMs in the context of a multi-topic question-answering task as a representative task that people may use LLMs for, e.g., in education and information retrieval. We find that frictions modulate over-reliance by driving down users' click rates while minimally affecting accuracy for those topics. Yet, frictions may have unintended effects. We find marked differences in users' click behaviors even on topics where frictions were not provisioned. Our contributions motivate further study of human-AI behavioral interaction to inform more effective and appropriate LLM use.

1 Introduction

There is a colloquial adage that "just because you can, does not mean you should." Large language models (LLMs) have seen unprecedented rates of use: OpenAI's ChatGPT had 100 million users within the first two months of release [41]. However, characterizing regimes of appropriate use is non-trivial: LLMs are general-purpose technologies with a plethora of use cases. LLMs may not be appropriate to deploy in all contexts as we have seen LLMs perform poorly at mathematics and arithmetic [10, 16], avoiding biased and hateful statements [19], and debugging code [44]. To better modulate when LLMs are used, we study the *selective* use of LLMs, thereby limiting access to their responses for specific queries, for specific users.

Mechanisms like reinforcement learning with human feedback [38] and direct preference optimization [40] take steps to steer LLM responses away from illegal, undesired, and toxic content. However, there are reasons beyond safety why it could be desirable to curb LLM use. In some contexts, discouraging the use of LLMs for particular users can yield economic or personal benefits. For example, encouraging students to solve problems on their own instead of accessing answers provided by LLMs may encourage a deeper understanding of and engagement with educational material [42, 53]. To prevent over-reliance on LLMs, dubbed "algorithm appreciation" [31], we advocate for thoughtful interactions with LLMs where users are vigilant about *when* they use these tools [52].

NeurIPS Workshop on Behavioral Machine Learning, 2024.

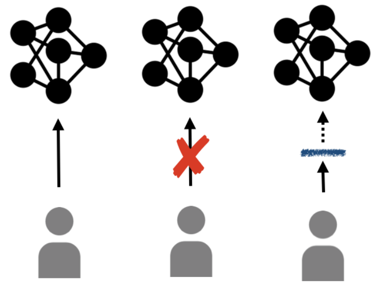

Figure 1: **Frictions permit continued model access, but require more effort to procure access**. Left: unrestricted access; middle: restricted access; right: frictioned access. We explore the use of selective frictions with respect to user expertise as a way to modulate the ease of model access across task instances.

To promote disuse, a spartan option is to restrict access to the LLM response entirely. For example, in deferral schemes, models abstain from providing predictions on specific task instances [33, 35]. As opposed to strict disuse wherein the LLM response is hidden, which could impair user freedoms, we consider selectively adding *friction* to an individual's experience with an LLM. Adapting a definition from Etzioni [15], we define a **friction** in the context of LLM assistance as:

> *A deliberate design element for increasing the time, effort, or cognitive load of accessing an AI-generated output by prompting conscious consideration of the task at hand.*

Also referred to as a nudge [51], microboundary [12], or cognitive forcing [13], frictions assuage algorithm appreciation and promote vigilant use of LLMs: users are encouraged to think twice before relying on an LLM. Similar to selecting when to abstain, introducing selective frictions can depend on model behavior, human expertise, or sociotechnical factors [3]. For instance, a friction could be selectively applied for a user who is relatively stronger than the LLM at some topic like mathematics.

We contribute a case study of the imposition of selective frictions, focusing on selectivity with respect to user expertise. Specifically, we consider a question-answer setting, reminiscent of "information-seeking", knowledge retrieval tasks that users may engage in with AI-based search summaries, e.g., Perplexity or Google Summaries. We extend the user interface, `Modiste`, from Bhatt et al. [4] to explore the imposition of an extra-click *selectively* (on only some topics, for some users) before the user can receive LLM assistance. We explore the impact of friction on users' click rates and attainable accuracy on multiple choice question from the popular and challenging NLP benchmark, MMLU [22] and other auxiliary measures, such as a user's confidence in their performance and that of the LLM. We observe marked behavior in users' click rates from the introduction of friction, providing initial evidence that frictioning LLM access can serve as one effective "lever" to design to modulate user experiences and help titrate overreliance. Yet, our study urges caution—to our surprise, we observe potential "spillover" effects where user behavior changes (i.e., reduced LLM engagements) even when no friction was imposed. Our study motivates further interdisciplinary human-centric work studying the interplay of human behavior, pragmatic inferences, and LLM predictions.

2 A Case Study in Selective Frictions

2.1 Task

We begin to explore the design and deployment of selective frictions for LLM experiences. Specifically, we consider assisted question-answering. Prior studies including Bhatt et al. [4], Mozannar et al. [36] have explored LLM assistance in answering multiple-choice questions from MMLU [22]; in particular, we build on the set-up and `Modiste` interface from Bhatt et al. [4], which supports rapid prototyping of user studies under various forms of assistance. Participants answer a total of 60 multiple-choice questions sampled from four topics of MMLU: US foreign policy, elementary mathematics, high school computer science, and high school biology.

In the baseline condition, participants can press a button to "query" the LLM and receive assistance on the current multiple-choice question (as shown in Figure 5), which then highlights one of the four multiple-choice options (as shown in Figure 6). As in many real-world settings, we selected a pool of questions where the model is not always correct, as it may not be in many real-world settings, as discussed in Appendix C.3.

2.2 Instantiation of Selective Friction

We propose a selective friction on top of this button by presenting the user with a second button requiring them to click again, indicating that they are certain that they want to see the model prediction. While the first button can be considered a friction in and of itself [6]; rather, we treat it as a baseline to compare *selective frictions*.

When to impose a friction? There are a variety of reasons for which a friction could benefit user subgroups, depending on the context. In our case study, we study one characteristic: user expertise. If one is already good at computer science, one may not benefit from access to the LLM prediction, particularly if the model has low accuracy.

To assess user expertise across the MMLU topics, we first have the user take a brief quiz (5 questions for each of the 4 topics). If they achieve higher performance than the LLM in a particular topic, then they will be presented with the friction for all questions of that topic in the "test" phase. If the user achieves the same expected topic performance as the model, when they indicate they want to see the model's prediction, there is no friction on access. We decide to friction a new question x if the following quantity is nonzero:

$$\texttt{friction}(x) = \mathbb{1}[\text{LLM}_{t(x)} > \text{User}_{t(x)}] \tag{1}$$

where $t(x)$ represents the topic of the query at hand, $\text{LLM}_{t(x)}$ represents the expected topic performance of the LLM, and $\text{User}_{t(x)}$ represents the expected topic performance based on the brief 5 question quiz. Details on the LLM performance for each topic are included in Appendix C.3.

How to present the friction? The friction takes the form of clicking a second button to view the LLM prediction. But what should this friction say? Small changes in wording can induce markedly different behavior in humans [20, 47]. To selectively friction by user expertise, we remind users of their expertise (relative to the model). We present performance as a frequency drawing on Lai and Tan [27] using the following template: `Do you really want to see the prediction? The AI model gets an average of X out of 10 questions correct on Math. Based on your warmup answers, we estimate that you get an average of Y out of 10 questions correct.` We present an example interface in Figure 3.

2.3 Participants

We recruit 100 participants from Prolific [39] in an institutionally-ethics reviewed study; participants are recruited from the US and required to speak English as a first language. Participants are randomly assigned to either the *selective-friction* or *baseline* condition ($N = 47$ and 53). In the friction condition, based on Eq. 1, 42 of 53 participants received friction for foreign policy, 49 for mathematics, 12 for computer science, and none for biology. We include more details in Appendix C.

2.4 Metrics

We focus on three metrics: (i) user accuracy over the questions for a given topic, (ii) click rate for questions in a given topic, which is the proportion of times that the user clicks to see the LLM prediction for M questions within a topic[1], and at the end of the study, (iii) the users' self-reported belief in their performance, as well as the LLM's performance, on each topic.

3 Results

Key Finding 1: Selective frictions can reduce click rates while maintaining accuracy. We analyze participant accuracy and click rates and conduct Ordinary Least Squares Regressions with

[1] In the frictioned setting, since the user technically needs to click twice before accessing the model, we only tally the second click.

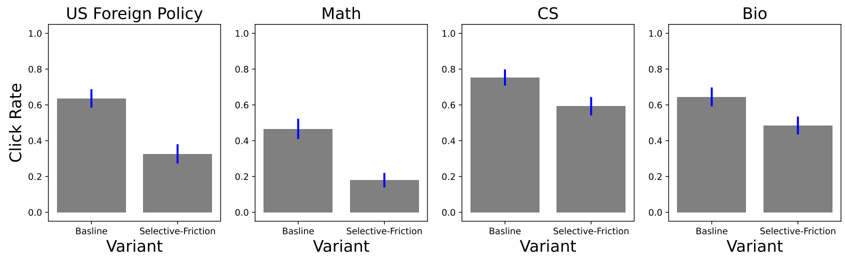

Figure 2: **Frictioning reduces clicks to see LLM predictions.** We measure the click rate for each user across topics. We find that, for all topics, click rates are statistically significantly reduced ($p < 0.05$) in the selective friction condition. Error bars indicate standard error over participants.

Benjamini-Hochberg correction, using a significance level of 0.05. We find that frictioning user experiences, in the way we have done here, indeed significantly lowers user click-through rates as shown in Figure 2 ($p < 0.05$). These results are encouraging, demonstrating friction may be one way to encourage users to solve problems independently. This finding is buttressed by minimal, not significant change in the users' accuracy, furthering the benefits of friction to users' critical thinking.

Key Finding 2: Frictions can induce unintentional spillover effects. However, to our surprise, we see that click rates drop for participants in the friction condition for biology—*even though no participants were frictioned specifically for biology*. This observation is important; frictioning users' experience in one region of the task space may influence users' decisions in other regions. We speculate why this may be happening, and encourage future work to empirically investigate this phenomenon further.

When a user is frictioned, we inform them of their own performance and that of the model (which, by definition of seeing the friction, is necessarily lower than the users'). The user may overgeneralize the lower model behavior on the frictioned topics to non-frictioned topics. We do observe a drop in users' predicted model accuracy in Table 1, across all topics. Alternatively, or additionally, frictions may encourage a user's own abilities and increase self-confidence in other questions. We correspondingly observe in Table 1 that users' self-confidence tends to increase in the frictioned setting.

4 Discussion

As users increasingly access powerful AI systems, buttressed by lightweight natural language interfaces, questions around how system designers can encourage and safeguard appropriate use grows more urgent. Our study demonstrates that small changes to user interfaces in the form of frictions can modulate user behavior, while preserving general user freedoms. We show that appropriately designed frictions can reduce user engagement and instances of over-reliance with minimal change in user accuracy. It is thus possible that frictions can serve as a critical tool in promoting the responsible use of LLMs. By incorporating barriers that encourage users to engage more critically with AI-generated content, policymakers can help ensure that LLMs are used thoughtfully and selectively, thus preserving and fostering users' impartiality and autonomy [2, 14, 46].

Here, we focused on adding hurdles along a user's path to engaging an LLM. Much work in the behavioral sciences has studied *positive interventions* to encourage particular kinds of behavior by "nudging" [47]. Next steps can explore nudges in our MMLU and other settings, as well as alternate mechanisms for instantiating selective frictions drawing on computational models of human behavior [8]. Further, while our selective frictioning design permits tailored user experiences, e.g., by expertise as we have shown, personalization of LLM experiences can come with risks [26].

While we find that our frictions can dampen excess engagement with an LLM when a user has appropriate expertise, we find that targeted frictions can have "spillover effects" wherein users' behavior changes even on topics where frictions were not added nor intended to be added. Our preliminary observations urge caution for designers of interventions around AI systems – human behavior is complex, and small changes in the realm of interaction may ripple into another. We are excited by future research at the intersection of AI and the behavioral sciences towards more effective "thought partners" [11].

Limitations

Our case study focuses on a single type of friction surrounding selective reminders with respect to user and model expertise; future work is needed to explore whether there are more effective frictions in terms of click rate modulation that may reduce spillover effects. Additionally, from our current study design, we cannot observe whether the user really needed and/or benefited from the LLM prediction. For example, some users clicked simply out of curiosity or to double-check their answers as noted in some post-survey responses (see Appendix D.1), which likely reduced the observed effect. As we only obtain the user's final prediction, future work might consider an alternative study design that asks the user for their answer before they see the LLM prediction, which may change the user's decision-making process. Another limitation of this work is our focus on a single dataset, MMLU. Since some of the tasks in question are quite challenging. We do not see any participants achieve high enough biology performance to be frictioned; many people may feel they need support from the LLM regardless. It is possible, as well, that our quiz does not obtain an adequate appraisal of participant expertise. We only evaluate participants on 5 questions per topic; as such, the expertise profile procured is necessarily an estimate – and we employ a necessarily reductive binary jurisdiction as to whether or not to apply a friction according to this coarse assesment of expertise. Future work is needed to explore alternative expertise elicitation schemes, and to understand whether click rate modulation and possible spillover effects in LLM experiences generalize to other settings and user populations.

References

[1] S. Alon-Barkat and M. Busuioc. Human–ai interactions in public sector decision making:“automation bias” and “selective adherence” to algorithmic advice. *Journal of Public Administration Research and Theory*, 33(1):153–169, 2023.

[2] V. Barletta, D. Caivano, D. Gigante, and A. Ragone. A rapid review of responsible AI frameworks: How to guide the development of ethical AI. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, pages 358–367. ACM, 2023. doi: 10.1145/3593434.3593478. URL `https://doi.org/10.1145/3593434.3593478`.

[3] U. Bhatt and H. Sargeant. When should algorithms resign? *arXiv preprint arXiv:2402.18326*, 2024.

[4] U. Bhatt, V. Chen, K. M. Collins, P. Kamalaruban, E. Kallina, A. Weller, and A. Talwalkar. Learning personalized decision support policies. *arXiv preprint arXiv:2304.06701*, 2023.

[5] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5 (CSCW1), apr 2021. doi: 10.1145/3449287. URL `https://doi.org/10.1145/3449287`.

[6] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[7] Z. Buçinca, S. Swaroop, A. E. Paluch, S. A. Murphy, and K. Z. Gajos. Towards optimizing human-centric objectives in ai-assisted decision-making with offline reinforcement learning. *arXiv preprint arXiv:2403.05911*, 2024.

[8] F. Callaway, M. Hardy, and T. L. Griffiths. Optimal nudging for cognitively bounded agents: A framework for modeling, predicting, and controlling the effects of choice architectures. *Psychological Review*, 2023.

[9] V. Chen, Q. V. Liao, J. Wortman Vaughan, and G. Bansal. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–32, 2023.

[10] K. M. Collins, A. Q. Jiang, S. Frieder, L. Wong, M. Zilka, U. Bhatt, T. Lukasiewicz, Y. Wu, J. B. Tenenbaum, W. Hart, et al. Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, 121(24):e2318124121, 2024.

[11] K. M. Collins, I. Sucholutsky, U. Bhatt, K. Chandra, L. Wong, M. Lee, C. E. Zhang, T. Zhi-Xuan, M. Ho, V. Mansinghka, et al. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10):1851–1863, 2024.

[12] A. L. Cox, S. J. Gould, M. E. Cecchinato, I. Iacovides, and I. Renfree. Design frictions for mindful interactions: The case for microboundaries. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, page 1389–1397, 2016. doi: 10.1145/2851581.2892410. URL `https://doi.org/10.1145/2851581.2892410`.

[13] P. Croskerry. Cognitive forcing strategies in clinical decisionmaking. *Annals of emergency medicine*, 41(1):110–120, 2003.

[14] V. Dignum. Responsible autonomy. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4698–4704. AAAI Press, 2017. ISBN 9780999241103.

[15] A. Etzioni. *A Socio-Economic Perspective on Friction*, chapter 3. Routledge, 2016.

[16] S. Frieder, L. Pinchetti, R. Griffiths, T. Salvatori, P. Lukasiewicz, T. Petersen, A. Chevalier, and J. Berner. Mathematical capabilities of chatgpt. *arXiv preprint*, 2023.

[17] Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

[18] B. Green and Y. Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[19] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

[20] D. Halpern. *Inside the nudge unit: How small changes can make a big difference*. Random House, 2016.

[21] L. Han. When the human is in the loop: Cost, effort and behavior. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2480–2480, 2020.

[22] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[23] D. Hummel and A. Maedche. How effective is nudging? a quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics*, 80:47–58, 2019. ISSN 2214-8043. doi: https://doi.org/10.1016/j.socec.2019.03.005. URL `https://www.sciencedirect.com/science/article/pii/S2214804318303999`.

[24] E. Jones, S. Sagawa, P. W. Koh, A. Kumar, and P. Liang. Selective classification can magnify disparities across groups. In *International Conference on Learning Representations*, 2020.

[25] B. Joshi, Z. Liu, S. Ramnath, A. Chan, Z. Tong, S. Nie, Q. Wang, Y. Choi, and X. Ren. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7103–7128, 2023.

[26] H. R. Kirk, B. Vidgen, P. Röttger, and S. A. Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, pages 1–10, 2024.

[27] V. Lai and C. Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 29–38, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287590. URL `https://doi.org/10.1145/3287560.3287590`.

[28] V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

[29] V. Lai, S. Carton, R. Bhatnagar, Q. V. Liao, Y. Zhang, and C. Tan. Human-AI collaboration via conditional delegation: A case study of content moderation. In *CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[30] Z. Li, Z. Lu, and M. Yin. Decoding ai's nudge: A unified framework to predict human behavior in ai-assisted decision making. *arXiv e-prints*, pages arXiv–2401, 2024.

[31] J. M. Logg, J. A. Minson, and D. A. Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.

[32] S. Ma, Y. Lei, X. Wang, C. Zheng, C. Shi, M. Yin, and X. Ma. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[33] D. Madras, T. Pitassi, and R. Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer, 2018.

[34] T. Mejtoft, S. Hale, and U. Söderström. Design friction. In *Proceedings of the 31st European Conference on Cognitive Ergonomics*, pages 41–44, 2019.

[35] H. Mozannar and D. Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087. PMLR, 2020.

[36] H. Mozannar, J. J. Lee, D. Wei, P. Sattigeri, S. Das, and D. Sontag. Effective human-ai teams via learned natural language rules and onboarding. *arXiv preprint arXiv:2311.01007*, 2023.

[37] H. Mozannar, J. Lee, D. Wei, P. Sattigeri, S. Das, and D. Sontag. Effective human-ai teams via learned natural language rules and onboarding. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[39] S. Palan and C. Schitter. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[40] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[41] R. Raman, S. Mandal, P. Das, and a. et. University students as early adopters of chatgpt: Innovation diffusion study. *Research Square preprint*, 2023. doi: https://doi.org/10.21203/rs.3.rs-2734142/v1.

[42] R. C. Schank, T. R. Berman, and K. A. Macpherson. Learning by doing. In *Instructional-design theories and models*, pages 161–181. Routledge, 2013.

[43] A. T. Schmidt and B. Engelen. The ethics of nudging: An overview. *Philosophy Compass*, 15(4):e12658, 2020. doi: https://doi.org/10.1111/phc3.12658. URL `https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12658`.

[44] D. Sobania, M. Briesch, C. Hanna, and J. Petke. An analysis of the automatic bug fixing performance of chatgpt. *arXiv preprint arXiv:2301.08653*, 2023.

[45] E. Sundin. Nudging and design friction: The impact on our decision making process. In *Conference in interaction technology and design*, page 39, 2021.

[46] C. Sunstein. *The Ethics of Influence: Government in the Age of Behavioral Science*. Cambridge University Press, 2016.

[47] R. H. Thaler and C. R. Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin, 2009.

[48] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.

[49] J. Von Kügelgen, A.-H. Karimi, U. Bhatt, I. Valera, A. Weller, and B. Schölkopf. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 9584–9594, 2022.

[50] Y. Wiener and R. El-Yaniv. Agnostic selective classification. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1665–1673. Curran Associates, Inc., 2011.

[51] J. Wilk. *Mind, Nature and the Emerging Science of Change: An Introduction to Metamorphology*, pages 71–87. Springer Netherlands, 1999. doi: 10.1007/978-94-017-2245-2_6. URL https://doi.org/10.1007/978-94-017-2245-2_6.

[52] J. Zerilli, U. Bhatt, and A. Weller. How transparency modulates trust in artificial intelligence. *Patterns*, page 100455, 2022.

[53] X. Zhu and H. A. Simon. Learning mathematics from examples and by doing. *Cognition and instruction*, 4(3):137–166, 1987.

A Ethics Statement and Potential Risks

As in the use of nudges from behavioral economics, there are critical conversations that warrant conversation around risks of selective frictions. While we take the stance as Thaler and Sunstein [47] that the choice to not adjust any access to users' experiences is still a choice, there are important questions around *who* is deciding when to impose frictions and on which user populations. While selective frictions could be one way to encourage critical thinking, as we begin to demonstrate here, without responsible design, they could negatively shape users' choice environments [46]. More pressingly, selective use of frictions may lead to disparate treatment of users, as some sub-populations of users may be frictioned on specific task instances more than others [24]. For instance, user expertise may be distributed unequally across a protected attribute; our friction framework would then disparately friction users increasing the *effort* required for some users to access the LLM output [21, 49]. Systems designers ought to be aware of such disparities and take the necessary precautions when deploying frictioned access to LLMs.

Table 1: **Frictioning induces minimal change in accuracy and may sway user belief in self- and model performance.** Per topic, we report average user accuracy and reported belief of expected self- and model performance (i.e., S-Belief and M-Belief respectively). Error bars indicate standard error over participants.

	US Foreign Policy			Math			CS			Bio		
Variant	Acc	S-Belief	M-Belief	Acc	S-Belief	M-Belief	Acc	S-Belief	M-Belief	Acc	S-Belief	M-Belief
Baseline	0.40 ± 0.04	0.35 ± 0.07	0.53 ± 0.06	0.67 ± 0.06	0.47 ± 0.08	0.63 ± 0.08	0.55 ± 0.04	0.25 ± 0.06	0.58 ± 0.07	0.67 ± 0.05	0.40 ± 0.07	0.60 ± 0.06
Friction	0.46 ± 0.05	0.47 ± 0.06	0.45 ± 0.06	0.66 ± 0.05	0.58 ± 0.07	0.57 ± 0.08	0.58 ± 0.04	0.36 ± 0.06	0.53 ± 0.07	0.67 ± 0.04	0.48 ± 0.06	0.55 ± 0.06

B Related Work

In this work, we focus on LLM-assisted user interactions and decision-making [18, 28]. Prior studies studying these contexts have shown the tendency for humans to overrely on AI support [9, 25, 48]. As such, recent works have considered adapting when AI support is provided to users: Ma et al. [32] fit a decision tree to offline users' decisions to decide when to show AI support to users, Buçinca et al. [7] use offline reinforcement learning to estimate if AI support would be helpful, and Bhatt et al. [4] employ online learning techniques to personalize a decision support policy to individuals. Relatedly, others have considered selectively delegating entire tasks to the AI model [17, 29, 33, 35, 50]. In many cases, it may not be possible (or desirable even if possible) to have an LLM make the final

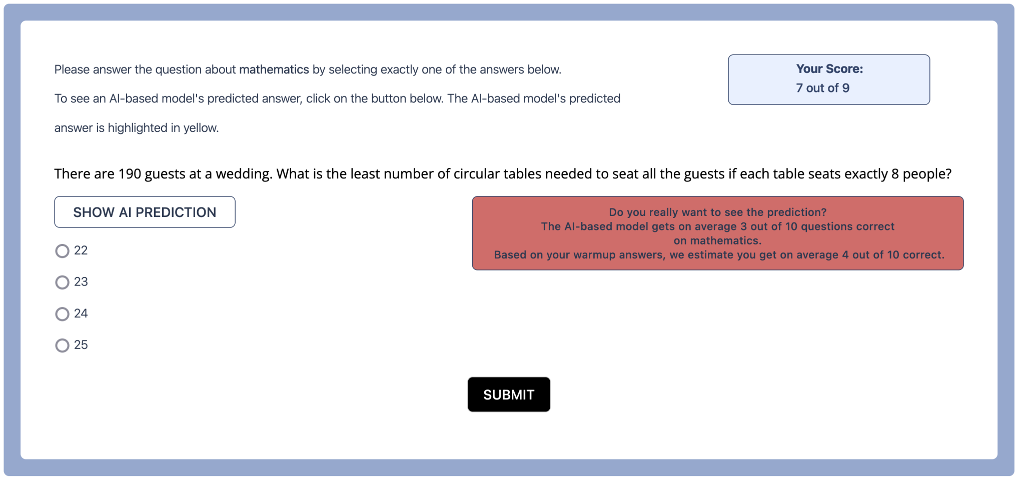

Figure 3: Interface for selective friction; here the user scored higher than the model in their pre-quiz on mathematics. If the user presses the "Show AI Prediction" button (first button) they are then presented with the red block (second button) which they are forced to click if they still want to see the prediction.

decision, and in others, it may not be justifiable to withhold user access to LLMs These constraints motivate our study of frictions, which permit continued model access but require more effort on the user's end to procure access.

Our work builds on the wealth of prior research into the design and effect of nudges on human behavior [23, 43, 47]. The notion of nudges is increasingly permeating machine learning, whether in the use of techniques from machine learning to design nudges [8] or nudging to support more appropriate use of AI systems [5, 30]. Relatedly, "microboundaries" [12], are small, intentional barriers integrated into user interfaces to promote more mindful interactions. Microboundaries can reduce the likelihood of users making errors or engaging in habitual, potentially harmful behaviors by interrupting their flow and requiring them to take an additional step before proceeding. There are several potential advantages of microboundaries (a la warning signals) as a type of design friction [1, 34, 45]. To our knowledge, we are the first work to explore selective frictioning of LLM use.

C Additional Details on Human Experiments

C.1 Participant Recruitment

We provide additional details on our user study. Participants receive the quiz for all conditions. The same questions are presented across both conditions, selected from three batches of 60 questions, as in [4]. The "test" phase involves 10 questions per topic. Participants are provided feedback as to whether they (and the model, if seen) are correct after each test trial. Feedback is not given in the quiz phase. Participants are paid at a base rate of \$9 per hour for an expected 30 minute experiment with an optional bonus up to \$10 per hour for correct answers; we apply the bonus to all participants. All data is anonymized, and participants provided informed consent before beginning the study.

C.2 Eliciting Perceived Self- and Model-Ability

At the end of the study, users are presented with a questionarre asking them to judge their own and the model's ability per topic. Specifically, we asked, for each topic: "Out of 100 questions on TOPIC, how many do you think *the AI* would get correct?" and "Out of 100 questions on TOPIC, how many do you think *you* could get correct (without the help of the AI-based model)?" For each question, users responded on an slider ranging from 0 to 100.

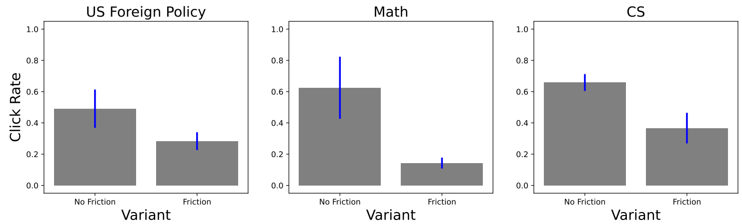

Figure 4: Comparing per-topic click rates people within the friction condition ($N = 53$) who did or did not see a friction. No one was frictioned for biology.

C.3 LLM Predictions

We use the same model predictions as in [4], which were sampled from InstructGPT3.5 `text-davinci-003` [38]; however, we randomly dampen model performance for the foreign policy and computer science topics such that the models achieve 30% and 60% performance on each, respectively. The model achieves approximately 30% and 90% performance on mathematics and biology. MMLU is challenging for humans [4, 22, 37]; our selective friction is triggered by user performance relative to the model's average performance on a topic – if the latter is too high, the friction will not be triggered. Accordingly, we expect with the dampening that we should have high trigger rates for mathematics and foreign policy, moderate rates for computer science, and low rates for biology – enabling us to study user click behavior across a range of model performances and settings wherein in some cases it is indeed rationale to rely on the model whereas in others, it may be disastrous for a user to regularly rely on the model (e.g., in elementary mathematics).

C.4 User Study Interface

We include example screenshots of the button interface and friction in Figures 5 and 3, respectively. When the user does click through to the LLM prediction, it is displayed as in 6, following Bhatt et al. [4].

D Additional Human Experiment Results

We include additional exploratory investigations into user behavior. We report average time spent for each topic in Table 2. We observe that the average time spent per problem appears to increase across topics.

We also decompose user behavior in Figure 4 within the frictioned condition according to whether the participant received a friction on that topic or not. Here, we can more clearly see that click rates decrease for participants who are explicitly frictioned. However, we caveat these results in that the friction intervention itself induces biases in the user samples across the groups. Participants only see a friction if they are necessarily better than the model; hence, users may already be inclined to click less often.

Table 2: Average time (seconds) per user per topic per question. We do not generally observe a significant difference in the amount of time spent as a result of the selective frictioning.

Variant	US Foreign Policy Time (sec)	Math Time (sec)	CS Time (sec)	Bio Time (sec)
Baseline	18.53 ± 1.59	23.23 ± 3.31	18.58 ± 2.50	22.80 ± 8.49
Friction Condition	23.29 ± 2.49	27.25 ± 3.15	18.85 ± 1.74	23.48 ± 3.72

D.1 Example User Responses

We asked participants in a post-survey questionarre what factors led them to click. We include a few exemplary responses below:

- “On questions where I didn’t immediately know the answer, I clicked to see the model predicted answer in case it would help me. On questions where I didn’t know the answer at all, I clicked to see the model predicted answer to help me know where to start.”
- “If I simply had no idea what the answer was to a question, I would click to see what the prediction would say and then decide if I wanted to go along with it or not. The other instance would be if I was stuck between two choices, I’d click the button to see if the prediction was the same answer as mine or not to have some sort of confirmation.”
- “not knowing anything about the topic or having any idea as to the right answer i clicked on the the ai button hoping it would help or know more than me. I also clicked it a couple times when i thought i knew the answer but wasnt 100% sure and if it chose the same as me i felt more confident”
- “Unsure if I had the right answer and not having enough knowledge or not having used knowledge of the subject for up to 30 years”
- “I used the prediction button if I felt unsure of the answer or if I wanted to feel more assured of my own answer.”
- “At first it was to help me with answering the question, then I realized the AI gave wrong answers as well, so for the ones I was sure of the answer I still clicked to see what it would show.It was very satisfactory to see I got it right when AI got it wrong, but very disappointing when AI gave me wrong answer when I didn’t know the correct answer”
- “If I was unsure of the correct answer (or second guessing myself), I checked the model prediction to see if the AI model aligned with what I was thinking”

When asked why they clicked, we noted that a few participants did report that they were simply curious:

- “In instances where I doubted my answer, I clicked to see the A.I model prediction. After getting an answer wrong I had an urge to click on the prediction. When I had no clue what to answer, I clicked the button. Sometimes, especially, for the math questions I clicked out of curiosity since I observed the A.I often got the answers wrong.”
- “I was mainly interested to see what it thought the answer was, independent if I thought I knew the answer or not. I had to look.”
- “I am not familiar with the terminology and am curious how AI responds.”

We also asked participants why they chose *not* to click:

- “I already 100% knew the answer so I didn’t wait to see what the model predicted answer was.”
- “I thought it would struggle to answer some of the more linguistically complicated questions correctly (like the ones that asked which of these is not true and 3 things are and 1 is not). I also thought for the most basic math problems (like things that were essentially a single computation), there wasn’t really a need.”
- “I didn’t click the button for answers where I was highly confident. (I suppose it probably wouldn’t have hurt to click it, but it would take a little pride out of it if the AI was correct too...)”
- “I like to challenge myself naturally, I’d prefer to make a good guess and be wrong to learn from it than to just look up the answer (or in this case, use AI) and immediately forget. I’d argue this is a case of disconnect between an internal effort and reward system.”
- “If I was fairly certain I already knew the answer, then I did not click it.”

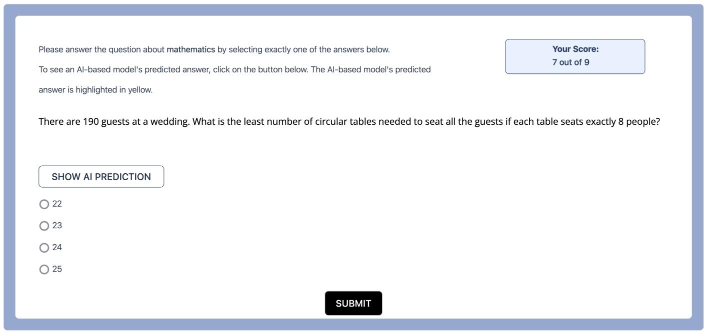

Figure 5: Example interface that the user is presented with for each MMLU question, where they have the option to click a button to query the AI.

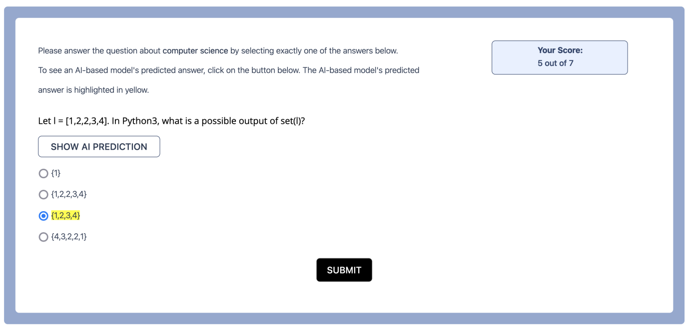

Figure 6: Example interface after the user has clicked the button to query the AI and the model prediction is shown via highlight.

- "The prompt [the friction] that came up was a major factor. It provided me with data that I was twice as likely to get it right than if I had not used it at all."
- "If I had a good idea or felt that I could answer quickly, I disregarded the AI suggestion since that would have wasted time."
- "[in the friction condition] Being told the AI would mostly get it wrong and confidence in my own ability to answer"
- "[in the friction condition] Being told the AI would mostly get it wrong and confidence in my own ability to answer"
- "I felt confident in my answer, and knowing that I was a better predictor in certain categories than the model"
- "If I thought I knew the answer to a question, I felt no reason to consult the AI. However, I did check the prediction when I had an answer, but I was unsure of it."

E Experiment Instructions

We include instructions presented in our user study in Figures 7, 8, and 9.

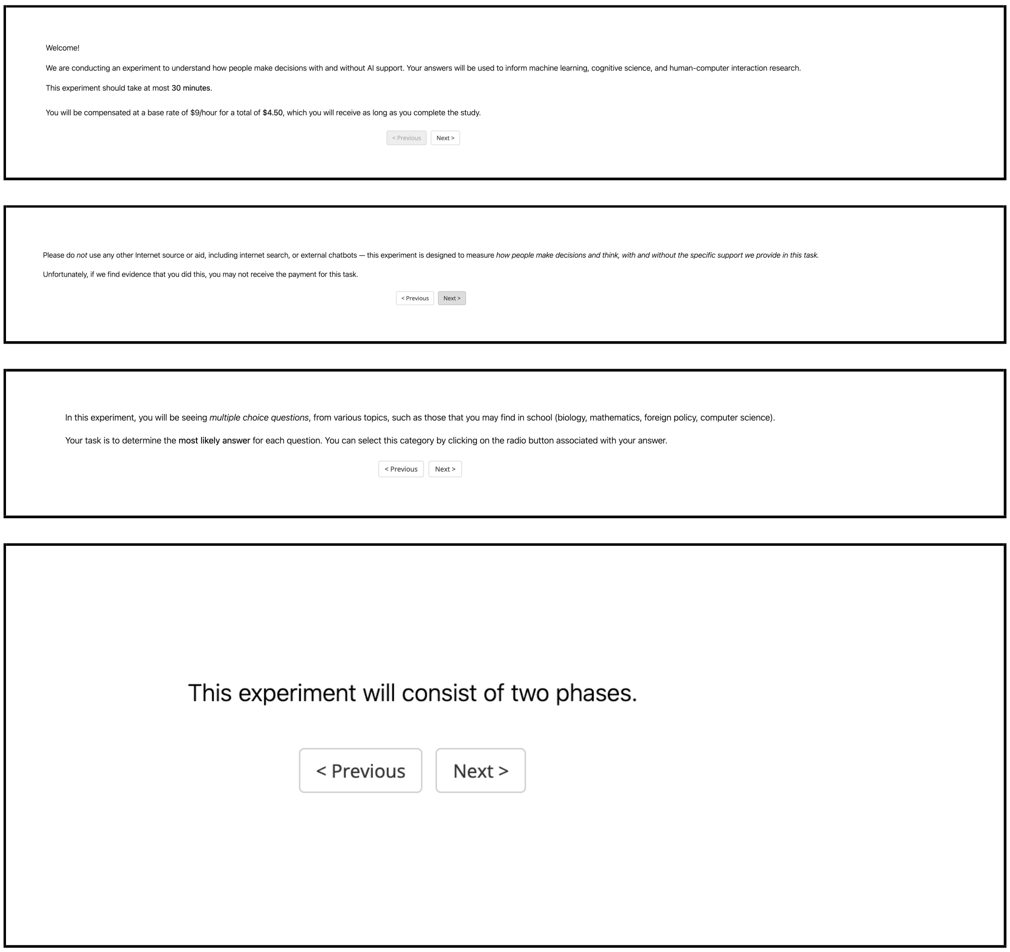

Figure 7: Experiment instructions.

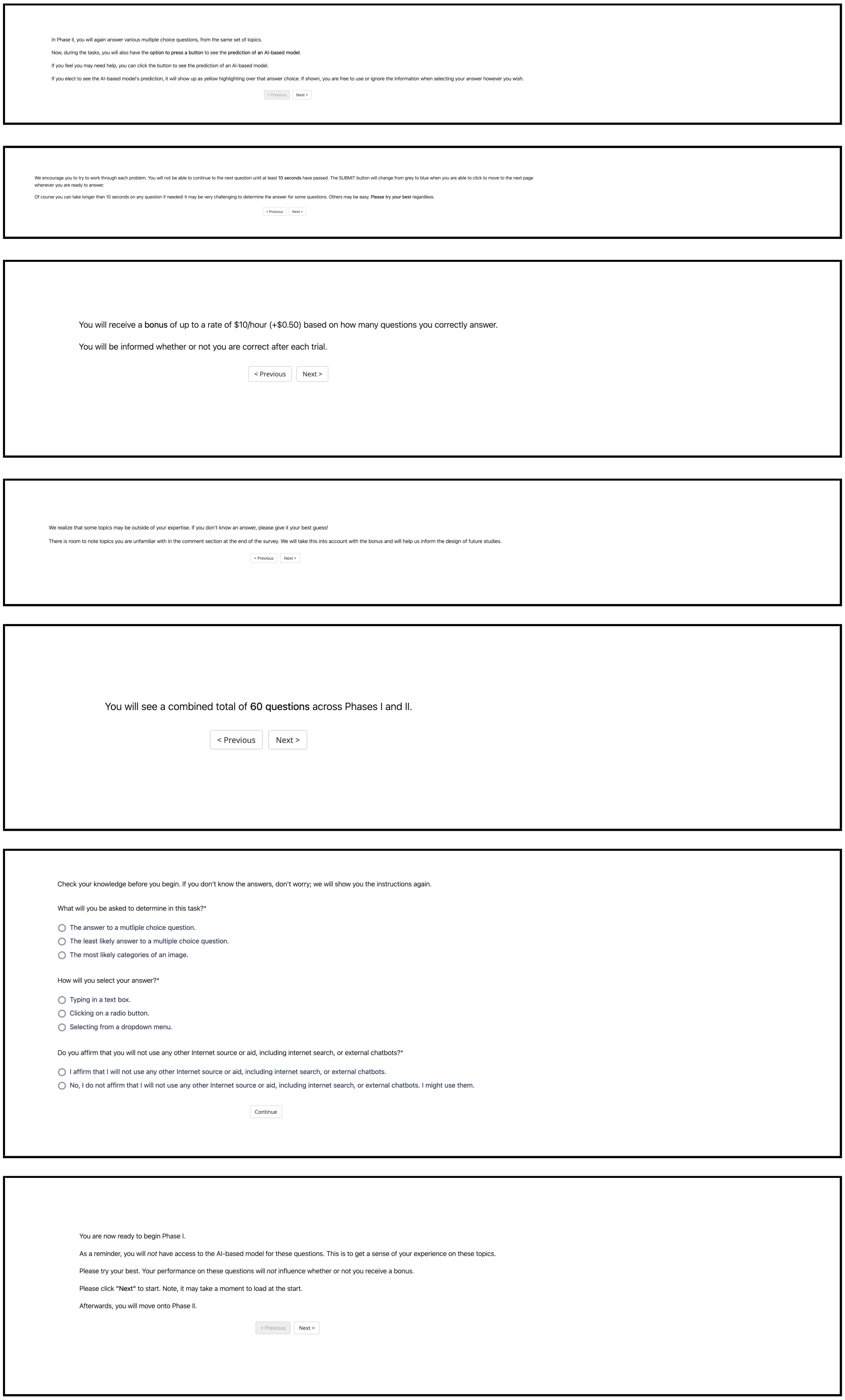

Figure 8: Experiment instructions (continued).

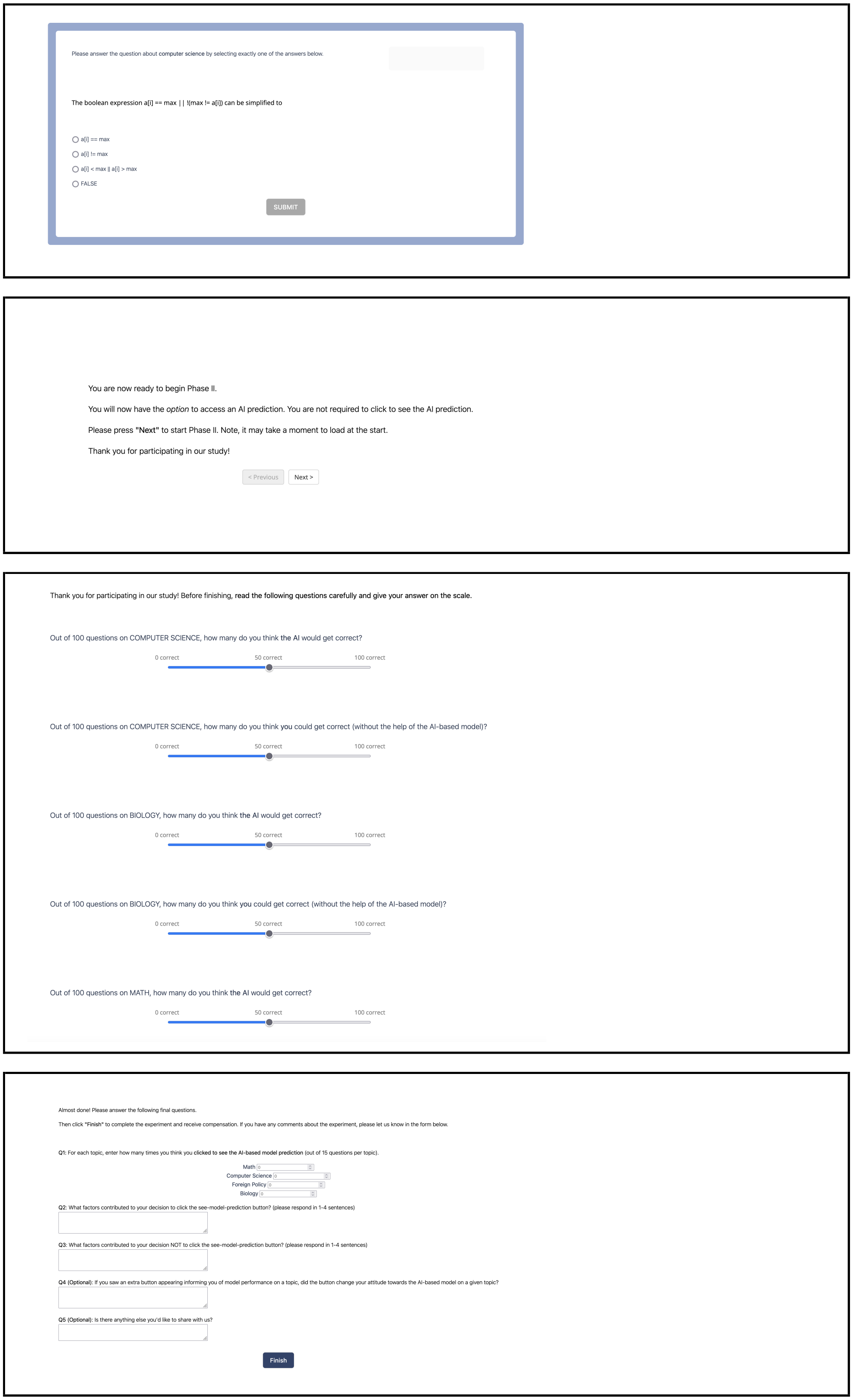

Figure 9: Experiment instructions (continued). The screen of Phase II is the official trial as presented in Figures 5, 6, and 3, respectively. The Phase I interface follows the same format, but no model access is permitted.