Proceedings of Machine Learning Research 273:146–173, 2025 iRAISE Workshop at AAAI 2025

Representational Alignment Supports Effective Teaching

Ilia Sucholutsky[*] IS3060@NYU.EDU
New York University

Katherine M. Collins[*] KMC61@CAM.AC.UK
University of Cambridge

Maya Malaviya[*] MAYA.MALAVIYA@NYU.EDU
New York University

Nori Jacoby KJ338@CORNELL.EDU
Cornell University

Weiyang Liu WL396@CAM.AC.UK
University of Cambridge

Theodore R. Sumers TED@ANTHROPIC.COM
Anthropic

Michalis Korakakis MK2008@CAM.AC.UK
University of Cambridge

Umang Bhatt UMANGBHATT@NYU.EDU
New York University

Mark Ho MARK.HO@NYU.EDU
New York University

Joshua B. Tenenbaum JBT@MIT.EDU
Massachusetts Institute of Technology

Brad Love B.LOVE@UCL.AC.UK
University College London

Zachary A. Pardos ZP@BERKELEY.EDU
University of California, Berkeley

Adrian Weller AW665@CAM.AC.UK
University of Cambridge

Thomas L. Griffiths TOMG@PRINCETON.EDU
Princeton University

Abstract

A good teacher should not only be knowledgeable, but should also be able to communicate in a way that the student understands – to share the student's representation of the world. In this work, we introduce a new controlled experimental setting, GRADE, to study pedagogy and representational alignment. We use GRADE through a series of machine-machine and machine-human teaching experiments to characterize a utility curve defining a relationship between representational alignment, teacher expertise, and student learning outcomes. We find that improved representational alignment with a student improves student learning outcomes (i.e., task accuracy), but that this effect is moderated by the size and representational diversity of the class being taught. We use these insights to design a preliminary classroom matching procedure, GRADE-Match, that optimizes the assignment of students to teachers. When designing machine teachers, our results suggest that it is important to focus not only on accuracy, but also on representational alignment with human learners.

[*] Equal contribution.

1. Introduction

The proliferation of digital education resources and AI systems has enabled human and machine teachers to reach potentially millions of students. For example, Massive Open Online Courses (MOOCs) promised to revolutionize education by having top educators record lectures in their domain of expertise and make course materials widely accessible online for many learners. However, this expert-first approach to online learning was not as effective and accessible as hoped for (Reich and Ruipérez-Valiente, 2019), with courses delivered by local teachers often showing better outcomes, in-person and online (Kelly, 2014). More recently, AI systems like ChatGPT have gained hundreds of millions of users, many of whom are using them, or the educational applications they power, to learn new subjects. While these systems can now outperform humans on some tasks (Strachan et al., 2023; Van Veen et al.; Thirunavukarasu et al., 2024), their internal representations are not often human-like (Fel et al., 2022; Muttenthaler et al., 2022), highlighting the distinction between domain expertise and the ability to map knowledge into human-understandable spaces. This tension is neither new nor unique to AI; professors can also be experts in their fields that struggle to communicate knowledge to students (Carter et al., 1987; Hinds et al., 2001). Yet many recent public education proposals explicitly focus on increasing teachers' domain expertise (e.g., Ontario, 2024), and much AI research continues to focus on improving the expertise of the agents being developed. Understanding further factors of effective teaching can help determine strategies for improving outcomes in classrooms.

We aim to bring together ideas from the burgeoning subfield of *representational alignment* (Sucholutsky et al., 2023b), machine teaching, and the cognitive science of pedagogy to shed light on further improvements for classrooms. We propose that 1) representational alignment between teachers and students, and 2) the size and diversity of the classroom, are both critical for determining the effectiveness of human and machine teaching (Figure 1B). To test this hypothesis, we design a simple modular student-teacher cognitive task environment called "Grid Manipulation of Representational Alignment and Domain Expertise" (GRADE) that enables the experimenter to independently control the teacher's expertise on the task, and the degree to which their representations of the task are similar to the student's (Figure 1A). Through simulations and a study where machines teach humans, we establish the relationship between teacher expertise, teacher-student alignment, and student performance. We find that representationally aligned teachers with a high error rate on the underlying task can outperform highly accurate but representationally misaligned teachers (Figure 1F). These results suggest that if a teacher adapts their representations to match the student, then the student's learning outcomes can significantly improve. We then extend our task from a teacher interacting with individual students to interacting with a class of representationally diverse students where the material they present is broadcasted to all students in the class (e.g., a lecture; see Figure 1D) and determine that class size and representational diversity moderate the effect of representational alignment on student outcomes (Figure 1G). Finally, we design a preliminary *classroom matching procedure*, GRADE-Match, that takes into account representational alignment, teacher expertise, and class size to optimize learning outcomes when assigning students to teachers (Figure 1E). We find that it outperforms both random assignment and MOOC-style assignment (Figure 1H). Our study emphasizes the importance of considering student-teacher representation alignment – not just teacher expertise – in pedagogical settings. This is especially important for designing AI thought partners that can think with us to help us grow (Collins et al., 2024) and tools that help personalize suggestions to individual students (Wang et al., 2024).

2. Related work

Learning Sciences. The extended learning sciences community has studied aspects of what makes for a good teacher or computer-based teaching system. The expertise or quality of the teacher with respect to excellence of schooling, certification, and a teacher's own test scores have been observed to positively affect student learning (Rice, 2003). More classroom-adaptive qualities, like a teacher's amount of experience in classrooms and teaching strategies employed (i.e., pedagogy) are also top contributing attributes (Rice, 2003). Closeness of representation to students with respect to demographic features has been shown to lead to more effective student performance (Dee, 2004), in part due to the role model effect, but also because teachers closer in these dimensions can serve as sociocultural interlocutors, helping translate the relevance of material to students (Egalite et al., 2015; Harfitt, 2018). Intelligent Tutoring Systems (Anderson et al., 1985), growing out of the cognitive and learning sciences, have been a consistently effective paradigm of computer-based teaching (Wang et al., 2023), primarily utilizing the pedagogy of mastery learning (Bloom, 1984). They adapt the amount of prescribed practice based on a representation of the student's level of mastery of the skill being worked on and provide procedural remediation in the problem-solving context. In a two-year, large-scale evaluation, a commercial ITS was found to be effective overall, but only demonstrated superior learning gains to standard classroom instruction in the second year. It was hypothesized that this may have been due to teachers needing to learn how best to align their classroom to the technology (Pane et al., 2014).

Machine teaching. Machine teaching aims to study the problem of teaching efficiency by characterizing such efficiency as the minimal number of effective data examples that is needed for a learner to learn some target concept. It has an intrinsic connection to optimal education (Zhu, 2015), curriculum learning (Liu et al., 2017; Korakakis and Vlachos, 2023) and optimal control (Lessard et al., 2019). Depending on the type of learner, machine teaching can be performed in a batch setting (Zhu, 2015; Zhu et al., 2018; Liu and Zhu, 2016) or an iterative setting (Liu et al., 2017, 2018, 2021; Qiu et al., 2023; Zhang et al., 2023). The batch teaching aims to find a training dataset of minimal size such that a learner can learn a target concept based on this minimal dataset. The iterative teaching seeks a sequence of data such that the learner can sequentially learn the target concept within minimal iterations. Complementary to these works, our findings indicate that, alongside the quality of examples that the teacher selects, it is also critical for both the teacher and the student to share similar representations.

Pragmatic communication. Successful communication rests on our ability to understand others' beliefs and intentions (Gweon, 2021; Vélez et al., 2023). Indeed, even young children are sensitive to others' knowledge and competence when teaching (Liszkowski et al., 2008; Bridgers et al., 2020) and learning (Bass et al., 2022; Csibra and Gergely, 2009; Bonawitz et al., 2011) from others. Inspired by Gricean pragmatics (Grice, 1975), recent computational models have formalized this process as recursive reasoning about others' latent mental states (Chen et al., 2022; Goodman and Frank, 2016; Shafto et al., 2014). Such pragmatic models have been used to study and facilitate human-AI interaction (Sumers et al., 2021, 2022; Lin et al., 2022; Andreas and Klein, 2016; Dale and Reiter, 1995; Fried et al., 2018; Wang et al., 2016, 2020; Ho et al., 2016; Zhi-Xuan et al., 2024; Liu et al., 2024). Crucially, however, when either party *fails* to accurately model the other's beliefs or perspective, human-human (Aboody et al., 2023; Sumers et al., 2023) and human-AI (Milli and Dragan, 2020; Sumers et al., 2022) communication can be significantly degraded. Our work adds to this literature by formalizing and analyzing the effect of *representational* misalignment on communication.

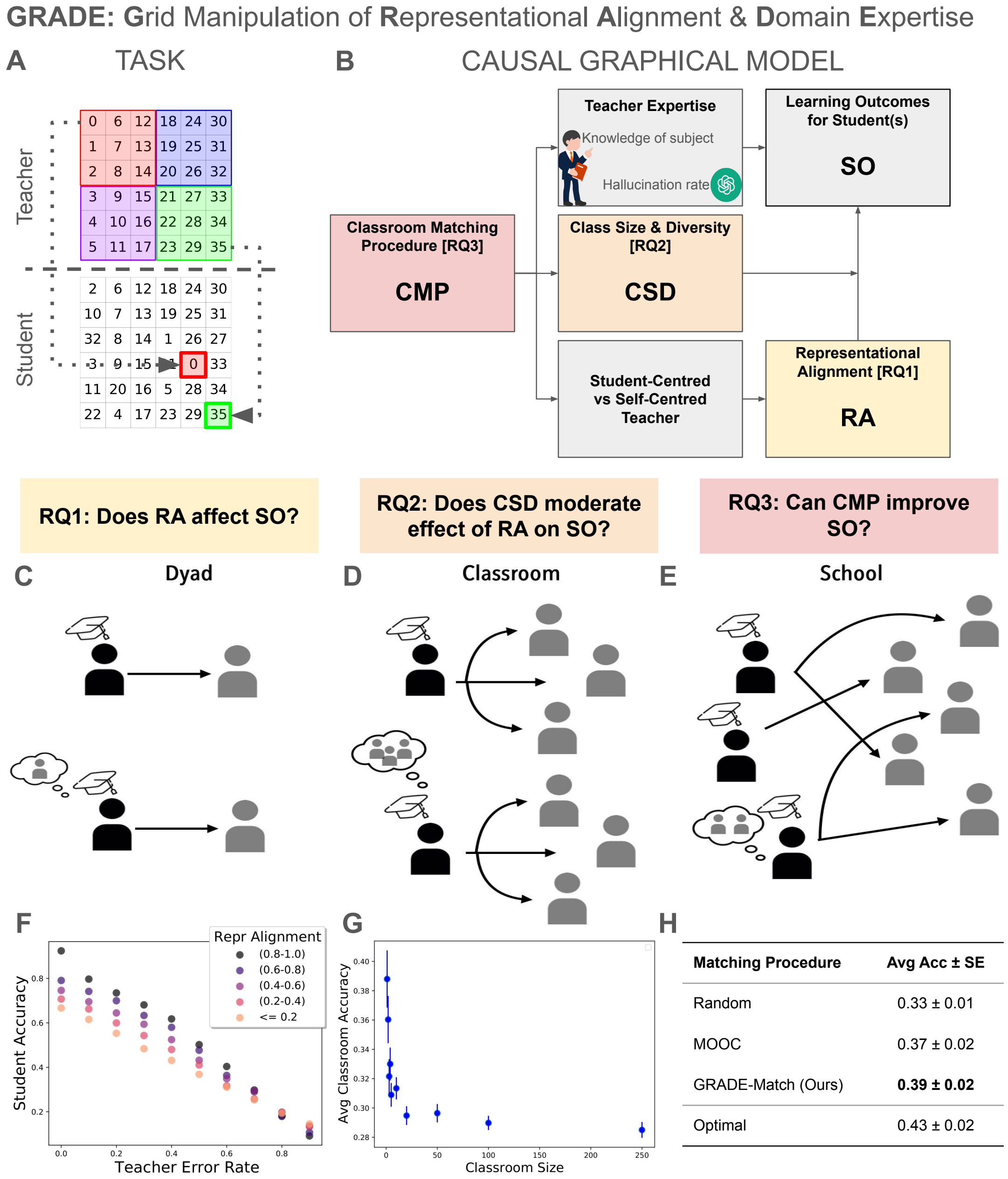
Representational alignment. Representational alignment (Sucholutsky et al., 2023b) offers a conceptual and grounded mathematical framework for characterizing teaching settings wherein two or more agents engage on some task. Already, ideas from representational alignment are providing new ways of thinking about machine learning efficiency (Sucholutsky et al., 2023a; Sucholutsky and Griffiths, 2023), value alignment (Rane et al., 2023; Wynn et al., 2023), disagreement (Oktar et al., 2023), and applications like human & machine translation and conversation (Niedermann et al., 2024; Huang et al., 2024). In this study, we show that representational alignment is a key dimension in predicting and optimizing student outcomes, with similar importance as the teacher's subject expertise.

3. GRADE: Grid Manipulation of Representational Alignment and Domain Expertise

We designed a new controlled task domain, called GRADE, where stimuli are arranged on a *grid* and labeled. Teachers know all the labels for the stimuli, though their expertise can be modulated by corrupting the true labels to get teachers with varying accuracy. Students do not yet know the labels but do see the stimuli. We focus on the setting where the teacher selects some labeled examples to reveal to the student. The arrangement of the stimuli on the grid might vary between the student and teacher, allowing us to manipulate representational alignment. We show an example of misaligned grids with two labeled examples chosen by the teacher in Figure 1A. GRADE lets us use any stimuli that can be arranged on a grid (e.g., based on pre-set features, as in the "salient-dinos" case in our human experiments, or even amortized embeddings). Here, we define representation alignment with respect to stimuli locations on the student and teacher grids; that is, we compute the Euclidean distance between pairwise swaps between stimuli. GRADE permits modularly-specifiable representation functions; we refer to Sucholutsky and Griffiths for a survey of a myriad of ways of measuring representation alignment. Additionally, we focus on a one-shot case where the teacher makes one round of selections. However, researchers can easily extend GRADE to multi-turn interactions. Appendix B contains a theoretical formalism of our setting.

4. RQ1: Does representational alignment affect teaching outcomes?

We begin to explore the relationship between representation alignment and student outcomes in two settings. We instantiate GRADE with two kinds of stimuli that can be arranged on an $N \times N$ grid with K underlying classes: **simple-features** (where each stimulus is only represented by its (x, y) coordinates; see Figure 7) and **salient-dinos** (stick figure images with features varying based on grid location). First, we simulate student-teacher interactions with simple 1-NN agents (see Appendix D.1) in the simple-features setting. We then generalize these findings with real human learners in both tasks. We include details of our teacher and student models, as well as our human experiment, in Appendix D.3.



Teacher

0	6	12	18	24	30
1	7	13	19	25	31
2	8	14	20	26	32
3	9	15	21	27	33
4	10	16	22	28	34
5	11	17	23	29	35

Student

2	6	12	18	24	30
10	7	13	19	25	31
32	8	14	1	26	27
3	9	15	21	0	33
11	20	16	5	28	34
22	4	17	23	29	35

Matching Procedure	Avg Acc ± SE
Random	0.33 ± 0.01
MOOC	0.37 ± 0.02
GRADE-Match (Ours)	**0.39 ± 0.02**
Optimal	0.43 ± 0.02

Figure 1: Overview. **A**: GRADE task; teacher and student receive misaligned grids (numbers only represent re-arranged elements, participants do not see them). Teacher is shown all labels (shown as colors) and reveals one per class to the student. Teacher's error rate is controlled by mislabeling their grid. **B**: Our hypothesized causal model. **C**: Dyadic interaction between a teacher and a student. "Student-centric" teachers infer the student's representations, making them fully aligned. **D**: "Classroom" setting where teacher broadcasts examples to all students (who have individual differences in representations); student-centric teachers jointly optimize over all students in the class. **E:** "School" setting where teachers are matched with students; each student is matched with a single teacher. **F**: Utility curve relating teacher error rate, representational alignment, and student accuracy in simulations. **G**: Average accuracy and standard errors in a student-centric class as a function of class size in simulations. **H**: Average accuracy and standard

errors across a school achieved by different matching procedures in simulations.

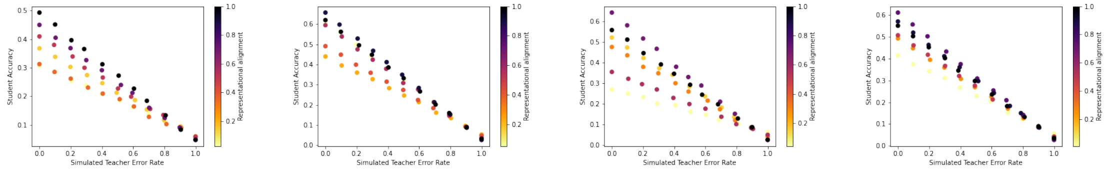


Figure 2: Relating teacher error rate and representational alignment between machine teachers and human students to student accuracy. From left to right: *simple-features* one class per quadrant; *salient-dinos* one per quadrant; *simple-features* one class per column (6); *salient-dinos* one class per column (7). Format follows Figure 1F.

4.1. Representational (mis)alignment in simulations

In Figure 1F, we trace out a relationship between student-teacher representational alignment, teacher error rate, and student accuracy. We uncover instances wherein students can achieve higher performance by learning from teachers who are *more erroneous* ("less expert") provided the teachers are representationally aligned with the students than comparatively more expert but representationally misaligned teachers, underscoring that it is not just the accuracy of a teacher that matters for student learning outcomes. For a fixed teacher error rate, higher representational alignment is always better for a student (provided the error rate is not too high). We uncover similar curves across grid sizes and the number of categories (see Appendix D.2).

4.2. Representational alignment of machine teachers and human students

We then generalize our findings through human experiments with $N = 480$ participants (see Appendix D.3). We construct a utility curve paralleling our simulations by post-hoc varying teacher error rate (see Appendix F.3). We find in Figure 2 that across both tasks (simple-features and salient-dinos), generally, higher representational alignment induces higher average student accuracy, and report correlations in Table 1. We find that even large increases in teacher error rate can be offset by increasing representational alignment (e.g., a teacher with error rate 0 and representational alignment of 0.3, has similar student outcomes as a teacher with error rate 0.4 and representational alignment 0.8). However, we note that the ordering of high representational alignment is less clear, particularly for the settings where each class corresponds to a column. We posit that people have a strong prior against classes being distributed as columns, and find that especially for the column conditions, participants would often label using strategies that did not correspond to nearest neighbor classification (e.g., several participants labeled in a way that corresponded to different types of tilings).

5. RQ2: How does class size and student representational diversity moderate the effect of representational alignment on student outcomes?

We have demonstrated that both a teacher's representational alignment and their accuracy matter for student outcomes. So far, our teachers have been self-centered; they use a single representation to select labeled examples. This approach suits machine teachers unable to adapt to specific students but does not capture capabilities of adaptable human or future machine teachers. Additionally, classroom teachers often address multiple students with differing representations, making example selection more complex. Here, we consider *student-centric* teachers who aim to maximize the average performance of a student pool by simulating likely student learning outcomes to various selections in an "inner loop" optimization (see Appendix E.6). While our earlier findings suggest student-centric teachers may enhance learning by becoming representationally aligned with students, we hypothesize that this effect will be moderated by the group's size and representational diversity, since teachers must optimize for all students simultaneously.

5.1. Setup

We extend GRADE to investigate classrooms of varying sizes. Because we sample students for each class from the same pool of representationally diverse students (Appendix E.1), increasing classroom size will generally increase representational diversity.

5.2. Results

We investigate the relationship between classroom size and student performance by sampling teachers with a range of error rates (between 0 and 0.5 in increments of 0.1) and classroom sizes (10 seeds per setting). Each student-centered teacher optimizes for their class through $T = 100$ inner loop iterations. We then marginalize over our sampled error rates to compute an expected average classroom accuracy per classroom size. We find that the performance of students in a classroom with a student-centric teacher is initially high (i.e., with a class of only a single student, the student-centric teacher would be equivalent to a fully representationally aligned teacher in the dyad setting) but falls off rapidly as a function of classroom size and then plateaus (as shown in Figure 1G).

6. RQ3: Can we match teachers and students to improve outcomes?

Given a "school" of teachers and students, how can we simultaneously group students into "classrooms" and determine which teacher to allot to which class? We begin to explore this question through a series of "classroom matching" experiments. We develop a *classroom matching procedure*, GRADE-Match, which given a pool of students and teachers, assigns groups of students to teachers based on our representational alignment-teacher utility curve. We emphasize though, that our analogy to "classrooms" and "schools" is explored in simulation with machines teaching machines; substantial future work is required to investigate the generalization of possible links between representational alignment, teacher error rate, and classroom properties in practice.

6.1. Setup

Student and teacher pools. We focus on our *simple-features* setting and extend our dyad (single teacher, single student) setting to simulated *pools* of teachers and students over our same 6×6 grid. We design two different pools of students and teachers (unstructured and structured). We include pool construction and generalization experiments to the *salient-dinos* setting in Appendix E.1 and E.5.

Matching procedures. We propose matching students using our utility curve to estimate their accuracy (**Grade-Match (Ours)**). We compute the representational alignment between a student and teacher and index into a bucketed version of the utility curve[1] that we construct in Section 3 using both the representational alignment and teacher's expected error rate (which we assume we have access to). The resulting metric is the student's expected performance under a specific teacher and classroom. We iterate over all teachers for each student and select the teacher who helps the student achieve the highest expected performance. We consider *three baselines*: (i) **Random** matching of students to teachers, (ii) **MOOC** which matches all students to the lowest error rate teacher, and (iii) **Optimal** wherein we use a brute force-search to match students to the highest attainable accuracy, giving an indication of the upper limit of performance that a matching algorithm could possibly achieve. Gaps between (ii) and (iii) further drive home the importance of going beyond teacher accuracy when pairing students to teachers.

6.2. Results

Our matching algorithm, which groups students to teachers based on their representational alignment and teacher error rate, generally outperforms random matching and, particularly for top-performing students, is better than having assigned the student to an expert (minimal error rate teacher; MOOC) who may be representationally distinct (see Figure 1H and Appendix Tables 2 and 3). This observation is intriguing – students may not achieve their full potential when paired with a representationally misaligned teacher, even if that teacher is an expert. We observe performance gains for our utility curve-based matching across both pool types. However, we do not yet attain optimal matching performance, perhaps due to a mismatch in our utility curve.

7. Discussion and limitations

Expertise on a task is not sufficient to be a good teacher; representational alignment matters too. Using a new controlled experimental paradigm (GRADE), we trace out a utility curve between teacher accuracy, teacher-student representational alignment, and student accuracy to characterize the crucial relationship between representational (mis)alignment and student learning outcomes. We put this utility curve to work to better match teachers to students based on representational alignment. Our work underscores the importance of teachers representing a diversity of students and arranging student-teacher groups to ensure there is at least one teacher that any student can effectively learn from. This motivates further investigation into representational alignment and its influence on pedagogical effectiveness in multiple learning settings, like teacher-student interactions and peer mentorship.

Yet, we emphasize that our work is a first step in the study of the relationship between representational alignment, teacher efficacy, and student-teacher matching. Our simulations always assume that students are 1-NN classifiers, which grossly undercuts the richness of human behavior. Further, our simulated students' representations are fixed; in practice, students adapt their representations over time. We also only consider single-turn, single-lesson settings, wherein students have no indication of the reliability of the teacher. We look forward to investigations that leverage and extend GRADE to go beyond our simple yet revealing initial setting.

[1] We recompute the curve by also averaging over samples of corrupted students as our first utility curve was constructed for dyad setting wherein the student's grid was never corrupted (see Appendix E.2).

Acknowledgments

We thank Alexis Ross, Dilip Arumugam, and Kartik Chandra for helpful conversations that informed this work. KMC gratefully acknowledges support from the Marshall Commission and the Cambridge Trust. AW acknowledges support from a Turing AI Fellowship under grant EP/V025279/1 and the Leverhulme Trust via CFI. This work was supported by a grant from NOMIS Foundation to TLG. MKH was supported by NSF Award #2434192.

References

Rosie Aboody, Joey Velez-Ginorio, Laurie R Santos, and Julian Jara-Ettinger. When naive pedagogy breaks down: Adults rationally decide how to teach, but misrepresent learners' beliefs. *Cognitive Science*, 47(3):e13257, 2023.

John R Anderson, C Franklin Boyle, and Brian J Reiser. Intelligent tutoring systems. *Science*, 228(4698):456–462, 1985.

Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. *arXiv preprint arXiv:1604.00562*, 2016.

Ilona Bass, Elizabeth Bonawitz, Daniel Hawthorne-Madell, Wai Keen Vong, Noah D Goodman, and Hyowon Gweon. The effects of information utility and teachers' knowledge on evaluations of under-informative pedagogy across development. *Cognition*, 222:104999, 2022.

Benjamin S Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16, 1984.

Elizabeth Bonawitz, Patrick Shafto, Hyowon Gweon, Noah D Goodman, Elizabeth Spelke, and Laura Schulz. The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, 120(3):322–330, 2011.

Sophie Bridgers, Julian Jara-Ettinger, and Hyowon Gweon. Young children consider the expected utility of others' learning to decide what to teach. *Nature Human Behaviour*, 4 (2):144–152, 2020.

Kathy Carter, Donna Sabers, Katherine Cushing, Stefinee Pinnegar, and David C. Berliner. Processing and using information about students: A study of expert, novice, and postulant teachers. *Teaching and Teacher Education*, 3(2):147–157, 1987. ISSN 0742-051X. doi: https://doi.org/10.1016/0742-051X(87)90015-1. URL https://www.sciencedirect.com/science/article/pii/0742051X87900151.

Alicia M Chen, Andrew Palacci, Natalia Vélez, Robert Hawkins, and Samuel J Gershman. A hierarchical Bayesian approach to adaptive teaching, Dec 2022.

Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, et al. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10):1851–1863, 2024.

Gergely Csibra and György Gergely. Natural pedagogy. *Trends in Cognitive Sciences*, 13 (4):148–153, 2009.

Robert Dale and Ehud Reiter. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263, 1995.

Thomas S Dee. Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics*, 86(1):195–210, 2004.

Anna J. Egalite, Brian Kisida, and Marcus A. Winters. Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review*, 45:44–52, 2015. ISSN 0272-7757. doi: https://doi.org/10.1016/j.econedurev.2015.01.007. URL https://www.sciencedirect.com/science/article/pii/S0272775715000084.

Thomas Fel, Ivan F Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems*, 35:9432–9446, 2022.

Christian Fischer, Zachary A. Pardos, Ryan Shaun Baker, Joseph Jay Williams, Padhraic Smyth, Renzhe Yu, Stefan Slater, Rachel Baker, and Mark Warschauer. Mining big data in education: Affordances and challenges. *Review of Research in Education*, 44 (1):130–160, 2020. doi: 10.3102/0091732X20903304. URL https://doi.org/10.3102/0091732X20903304.

Michael Frank. Modeling the dynamics of classroom education using teaching games. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 31, 2018.

Noah D Goodman and Michael C Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829, 2016.

Herbert P Grice. Logic and conversation. In *Speech Acts*, pages 41–58. Brill, 1975.

Hyowon Gweon. Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, 25(10):896–910, 2021.

G Harfitt. The role of the community in teacher preparation: Exploring a different pathway to becoming a teacher. *Front. Educ. 3: 64. doi: 10.3389/feduc*, 2018.

Pamela J Hinds, Michael Patterson, and Jeffrey Pfeffer. Bothered by abstraction: the effect of expertise on knowledge transfer and subsequent novice performance. *Journal of Applied Psychology*, 86(6):1232, 2001.

Mark K Ho, Michael Littman, James MacGlashan, Fiery Cushman, and Joseph L Austerweil. Showing versus doing: Teaching by demonstration. *Advances in Neural Information Processing Systems*, 29, 2016.

Dun-Ming Huang, Pol Van Rijn, Ilia Sucholutsky, Raja Marjieh, and Nori Jacoby. Characterizing similarities and divergences in conversational tones in humans and llms by sampling with people. *arXiv preprint arXiv:2406.04278*, 2024.

Andrew P Kelly. Disruptor, distracter, or what? a policymaker's guide to massive open online courses (MOOCs). *Bellwether Education Partners*, 2014.

Michalis Korakakis and Andreas Vlachos. Improving the robustness of NLI models with minimax training. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 14322–14339, 2023. doi: 10.18653/V1/2023.ACL-LONG.801. URL https://doi.org/10.18653/v1/2023.acl-long.801.

Laurent Lessard, Xuezhou Zhang, and Xiaojin Zhu. An optimal control approach to sequential machine teaching. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2495–2503, 2019.

Jessy Lin, Daniel Fried, Dan Klein, and Anca Dragan. Inferring rewards from language in context. *arXiv preprint arXiv:2204.02515*, 2022.

Ulf Liszkowski, Malinda Carpenter, and Michael Tomasello. Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition*, 108 (3):732–739, 2008.

Ji Liu and Xiaojin Zhu. The teaching dimension of linear learners. *Journal of Machine Learning Research*, 17(162):1–25, 2016.

Ryan Liu, Jiayi Geng, Joshua C Peterson, Ilia Sucholutsky, and Thomas L Griffiths. Large language models assume people are more rational than we really are. *arXiv preprint arXiv:2406.17055*, 2024.

Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B Smith, James M Rehg, and Le Song. Iterative machine teaching. In *International Conference on Machine Learning*, pages 2149–2158. PMLR, 2017.

Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, James Rehg, and Le Song. Towards blackbox iterative machine teaching. In *International Conference on Machine Learning*, pages 3141–3149, 2018.

Weiyang Liu, Zhen Liu, Hanchen Wang, Liam Paull, Bernhard Schölkopf, and Adrian Weller. Iterative teaching by label synthesis. *Advances in Neural Information Processing Systems*, 34:21681–21695, 2021.

Yuzhe Ma, Robert Nowak, Philippe Rigollet, Xuezhou Zhang, and Xiaojin Zhu. Teacher Improves Learning by Selecting a Training Subset. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1366–1375, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/ma18a.html.

Maya Malaviya, Ilia Sucholutsky, Kerem Oktar, and Thomas L Griffiths. Can humans do less-than-one-shot learning? In *44th Annual Meeting of the Cognitive Science Society: Cognitive Diversity, CogSci 2022*, 2022.

Smitha Milli and Anca D Dragan. Literal or pedagogic human? analyzing human model misspecification in objective learning. In *Uncertainty in Artificial Intelligence*, pages 925–934. PMLR, 2020.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. *arXiv preprint arXiv:2211.01201*, 2022.

Jakob Niedermann, Ilia Sucholutsky, Raja Marjieh, Elif Celen, Thomas L Griffiths, Nori Jacoby, and Pol van Rijn. Studying the Effect of Globalization on Color Perception using Multilingual Online Recruitment and Large Language Models, Feb 2024. URL osf.io/preprints/psyarxiv/3jvxw.

Kerem Oktar, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. Dimensions of disagreement: Unpacking divergence and misalignment in cognitive science and artificial intelligence, 2023.

Ontario, May 2024. URL https://news.ontario.ca/en/backgrounder/1004649/modern-relevant-and-skills-focused-a-stronger-ontario-high-school-diploma.

Stefan Palan and Christian Schitter. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

John F Pane, Beth Ann Griffin, Daniel F McCaffrey, and Rita Karam. Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis*, 36(2): 127–144, 2014.

Zeju Qiu, Weiyang Liu, Tim Z Xiao, Zhen Liu, Umang Bhatt, Yucen Luo, Adrian Weller, and Bernhard Schölkopf. Iterative teaching by data hallucination. In *International Conference on Artificial Intelligence and Statistics*, pages 9892–9913, 2023.

Sunayana Rane, Mark Ho, Ilia Sucholutsky, and Thomas L Griffiths. Concept alignment as a prerequisite for value alignment. *arXiv preprint arXiv:2310.20059*, 2023.

Justin Reich. *Failure to disrupt: Why technology alone can't transform education*. Harvard University Press, 2020.

Justin Reich and José A Ruipérez-Valiente. The MOOC pivot. *Science*, 363(6423):130–131, 2019.

Jennifer King Rice. *Teacher quality: Understanding the effectiveness of teacher attributes.* ERIC, 2003.

Patrick Shafto, Noah D Goodman, and Thomas L Griffiths. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, 71: 55–89, 2014.

James Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Alessandro Rufo, Guido Manzi, Michael Graziano, and Cristina Becchio. Testing theory of mind in GPT models and humans. 2023.

Ilia Sucholutsky and Thomas L Griffiths. Alignment with human representations supports robust few-shot learning. *NeurIPS*, 2023.

Ilia Sucholutsky, Ruairidh M Battleday, Katherine M Collins, Raja Marjieh, Joshua Peterson, Pulkit Singh, Umang Bhatt, Nori Jacoby, Adrian Weller, and Thomas L Griffiths. On the informativeness of supervision signals. In *Uncertainty in Artificial Intelligence*, pages 2036–2046. PMLR, 2023a.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi, Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment, 2023b.

Theodore Sumers, Robert Hawkins, Mark K Ho, Tom Griffiths, and Dylan Hadfield-Menell. How to talk so AI will learn: Instructions, descriptions, and autonomy. *Advances in Neural Information Processing Systems*, 35:34762–34775, 2022.

Theodore R Sumers, Mark K Ho, Robert D Hawkins, Karthik Narasimhan, and Thomas L Griffiths. Learning rewards from linguistic feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6002–6010, 2021.

Theodore R Sumers, Mark K Ho, Robert D Hawkins, and Thomas L Griffiths. Show or Tell? Exploring when (and why) teaching with language outperforms demonstration. *Cognition*, 232:105326, 2023.

Arun James Thirunavukarasu, Shathar Mahmood, Andrew Malem, William Paul Foster, Rohan Sanghera, Refaat Hassan, Sean Zhou, Shiao Wei Wong, Yee Ling Wong, Yu Jeat Chong, Abdullah Shakeel, Yin-Hsi Chang, Benjamin Kye Jyn Tan, Nikhil Jain, Ting Fang Tan, Saaeha Rauz, Daniel Shu Wei Ting, and Darren Shu Jeng Ting. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study. *PLOS Digital Health*, 3(4):1–16, 04 2024. doi: 10.1371/journal.pdig.0000341. URL https://doi.org/10.1371/journal.pdig.0000341.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. Clinical text summarization: Adapting large language models can outperform human experts. *Research Square*.

Natalia Vélez, Alicia M Chen, Taylor Burke, Fiery A Cushman, and Samuel J Gershman. Teachers recruit mentalizing regions to represent learners' beliefs. *Proceedings of the National Academy of Sciences*, 120(22):e2215015120, 2023.

Huanhuan Wang, Ahmed Tlili, Ronghuai Huang, Zhenyu Cai, Min Li, Zui Cheng, Dong Yang, Mengti Li, Xixian Zhu, and Cheng Fei. Examining the applications of intelligent tutoring systems in real educational contexts: A systematic literature review from the social experiment perspective. *Education and Information Technologies*, 28(7):9113–9148, 2023.

Pei Wang, Junqi Wang, Pushpi Paranamana, and Patrick Shafto. A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems*, 33: 17582–17593, 2020.

Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, and Dora Demszky. Tutor copilot: A human-ai approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*, 2024.

Sida I Wang, Percy Liang, and Christopher D Manning. Learning language games through interaction. *arXiv preprint arXiv:1606.02447*, 2016.

Andrea Wynn, Ilia Sucholutsky, and Thomas L Griffiths. Learning human-like representations to enable learning human values. *arXiv preprint arXiv:2312.14106*, 2023.

Teresa Yeo, Parameswaran Kamalaruban, Adish Singla, Arpit Merchant, Thibault Asselborn, Louis Faucon, Pierre Dillenbourg, and Volkan Cevher. Iterative classroom teaching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5684–5692, 2019.

Chen Zhang, Xiaofeng Cao, Weiyang Liu, Ivor Tsang, and James Kwok. Nonparametric iterative machine teaching. In *International Conference on Machine Learning*, pages 40851–40870, 2023.

Tan Zhi-Xuan, Lance Ying, Vikash Mansinghka, and Joshua B. Tenenbaum. Pragmatic instruction following and goal assistance via cooperative language-guided inverse planning, 2024.

Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. An overview of machine teaching. *arXiv preprint arXiv:1801.05927*, 2018.

Contents

Appendix A. Broader impact and societal risks

As we discuss in Section 7, our work portends broader implications for the design of machine-human teaching setups where machines are intentionally built with representation alignment in mind, as well as representation diversity to safeguard against threats to inclusivity. It is possible that our simulations could support interventions in real classrooms, e.g., informing classroom size decisions drawing on measures of the representational diversity of a classroom pool and the expertise of the teacher. However, we heed caution in over-generalizing our results to settings where real student experiences and learning potential is at stake. Broad-brush application of AI systems in education has not been met with universal success (Reich, 2020) and inappropriately incorporated can have unintended impacts on student success (Fischer et al., 2020).

Appendix B. Theoretical formulation underlying GRADE

We offer a deeper theoretical formalism for our setting. Figure 3 shows a schematic of our teaching and representation alignment framework. Consider a space X of stimuli. We consider the case where the teacher tries to teach the students some function $f : X \to C$. We illustrate a simple case in Figure 3 wherein C is a binary classification $C = \{0, 1\}$ dividing X into two regions ($C = 0$ and $C = 1$ are represented in Fig. 3 in light and dark gray, respectively). The teacher observes label function $f' : X \to C$, which may be different from f.

The teacher chooses n points from the space $x_1, x_2, \ldots, x_n \in X$ and assigns labels to the points l_i. The teacher materials can be represented by the labeled points : $L_0 = (x_i, l_i)_{i=1,\ldots,n}$. To represent the fact that students' representations may differ from that of the teacher, we assume that the student s has a space Y_s that corresponds to the student's representations. Note that each student is part of some classroom or population of students $s \in \mathcal{S}$.

Next, we assume there is some transformation $T_s : X \to Y_s$. We assume that the function T_s is also selected from some parametric function $T_s \sim \mathcal{T}$. The student s observes stimuli

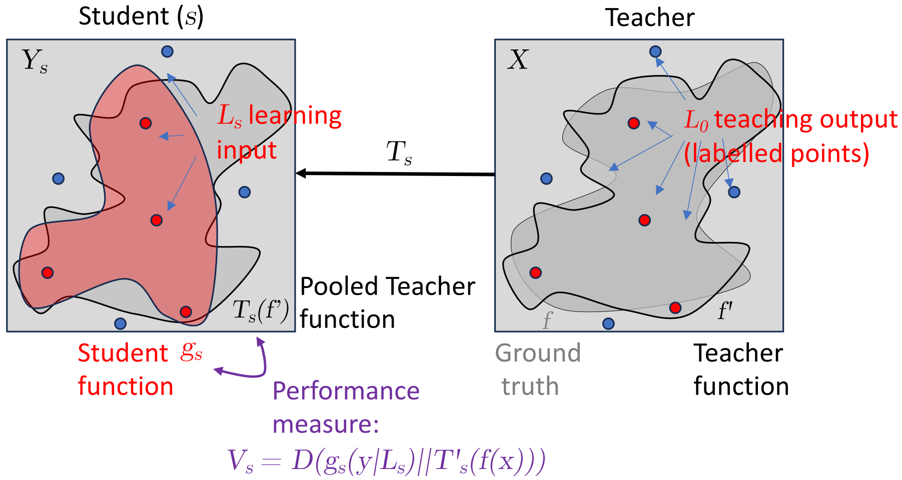


Figure 3: Schematic of teaching and representational alignment. Teachers and students have distinct representational spaces (X, Y_s) with some mapping between them (T_s). There is a true label function (f) that can be projected onto both the teacher and student spaces, but a teacher may not perfectly know this true label function and have their own, diverging label function (f'). The teacher designs curricular materials (L_0; a set of examples paired with labels) that are projected to each student's space (L_s), where each student uses them to learn a label function (g_s). Each student's performance (V_s) is then measured as the divergence between the learned label function and the hidden true label function $(T_s'(f))$.

presented by the teacher $y_i = T_s(x_i)$ and labels l_i. The student learning input (i.e., the teaching materials mapped into the student's space) is thus $L_s = (T_s(x_i), l_i)_{i=1,...,n}$. From that, the student infers the labeling for the rest of the space, which can be represented as the learning function $g_s(y|L_s)$. The classification performance of the student is tested over additional test points where the expected performance of the student is $V_s = D(g_s(y|L_s)||T_s(f(x)))$. Here, D represents some distance measure.

Appendix C. Compute details

All experiments were run on an 8-core, 16 GB memory laptop. Experiments were run exclusively on CPUs and were all runnable within at most three hours. Our experiments are reproducible and all the implementations for all computational experiments will be made available open-source upon publication.

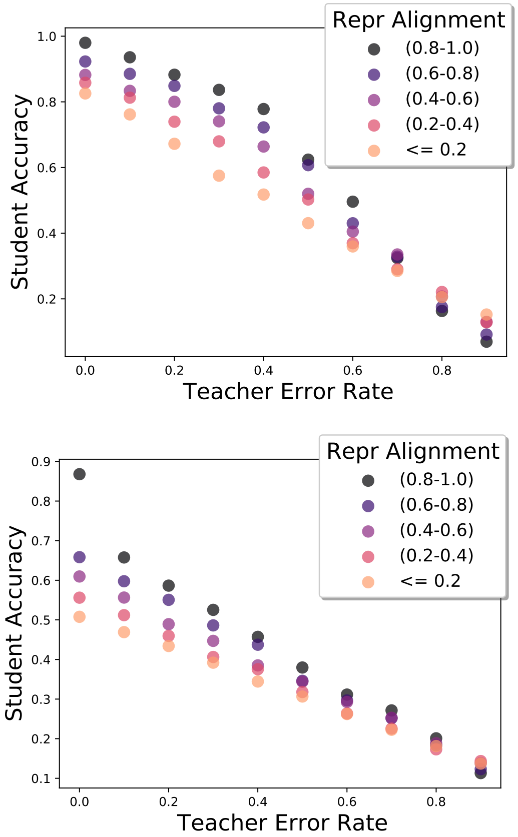
Appendix D. Additional details on task setup for the single teacher-single student setting

D.1. Student and teacher models

We instantiate our student (g_s) as a 1-nearest neighbor (1-NN) classifier, who takes as input the teacher's revealed examples (L_s) and classifies each of the unlabeled points. Student performance (V_s) is computed as the accuracy of their classifications over the unlabeled points. The teacher chooses K' points intended to maximally help the student (whom the teacher "knows" is a 1-NN classifier) to achieve high accuracy on the remaining points. We assume the teacher has access to labels for all cells; however, the "erroneous" teacher with some probability assumes the wrong label on a cell (i.e., f' is different from f, which can ripple into their selections accordingly). The teacher computes the centroid of each class (using its own believed labels f', which may have errors) and selects one example per class to reveal to the student. The teacher reveals its believed labels to the student for the selected points. While we limit our analyses to assuming 1-NN classifiers, in principle, GRADE can be used with any student model and future research should focus on developing and analyzing more accurate models of human students.

D.2. Constructing the dyadic utility curve

To construct our utility curve in Section 3, we sweep over a range of possible teacher error rate parameterizations (from 0 to 0.9 in increments of 0.1) and representation corruption levels (from 0 to 1.0 in increments of 0.01). We always use the same "student" and corrupt teachers over the respective student grid. We compute the representation alignment between the student and corrupted teacher; as the pairwise swaps ("corruptions") are randomly made over a fraction of the grid parameterized by the corruption level, we bucketize the resulting observed representation alignment between student and teacher. We then sample 10 different seeds of selections for each teacher and average student performance. We repeat the same sweeps over teacher parameterizations for our two labeling schemes: grids wherein each column corresponds to one label (N labels for an $N \times N$ grid), and one where each



(a)

(b)

Figure 4: Utility curves on a 6 × 6 grid for different label structures. **(Left:)** 4 class underlying label-per-quadrant; **(Right:)** 6 class underlying label-per-column.

quadrant corresponds to one label (four labels). We average the resulting utility curves across label types.

Impact of underlying label structure We depict the separate utility curves in Figure 4. Notably, we observe different utility curves for different label structures. While there are some minor rank swaps between teachers across the structures, we see high Spearman rank correlation ($\rho = 0.994, p << 10e - 48$) between the two settings underscoring general consistency in teacher orderings.

D.3. Additional details on experiments with machines teaching humans

D.3.1. Participants.

We recruit 480 participants from Prolific (Palan and Schitter, 2018). We filtered the participant pool by country of residence (United States), number of completed studies (> 100), and approval rate ($> 95\%$). Participants gave informed consent under an approved IRB protocol.

D.3.2. Task.

We design a task for our machines to teach humans about categories, in which participants see a grid of stimuli; for each cell in the grid, there is an underlying true category. Our simulated teacher model selects labels based on these underlying categories, and participants see these labels with the grid. Participants must then categorize all the unlabeled stimuli on the grid using the teacher's labels. We do not inform participants of the number of examples per category. We investigate two structures of categories: one class per column of the grid ("cols") and one class per quadrant of the grid ("quad"). Note that these categories induce labeling functions that the students *should* be able to learn; they are tractable (column structure and block structure). There were two different stimuli sets. The first (*simple-features*) is the closest analog to our simulated experiments, in which participants saw a 6×6 grid with blank cells, so the features are completely expressed via the coordinates of the grid. The second (*salient-dinos*) is a more rich set of stimuli, wherein participants see a 7×7 grid of dinosaur ("dino") images from Malaviya et al. (2022). Dino stimuli were defined by nine different features (e.g., body length, neck length, neck angle) and organized on the grid by two principle components of those underlying features. For a visualization of the participant's view, see Figure 7. For each condition, different teachers were generated from our model, sampling across varying levels of alignment. This structure leads to 24 different conditions (2 stimuli sets × 2 category structures × 6 teacher alignment levels) for which we collect 20 participants each.

D.3.3. Models and example selection.

We employ the same model types as in our simulations. Teachers are self-centered and assume that students are 1-NN classifiers[2]. In both settings, we assume the representations of teachers and students can be expressed through their two-dimensional grid locations. For

[2] We acknowledge such an assumption is highly simplistic for students and encourage future work to explore alternate models of students.

the simple-grid setting, there are no features for the human to use for their categorization beyond grid cell location; and in the *salient-dinos* setting, features were defined by two principal components (which we can use as grid coordinates). We again induce representation misalignment between teacher and student by shuffling the stimuli on the grid. We sample a set of teachers spanning a range of representational alignments. We select a single set of points for each teacher, assuming the teacher has perfect accuracy. We explore alternate labeling functions to simulate alternate teacher error rates post-hoc (see Appendix F).

D.3.4. Additional results on machines teaching humans

We present the correlations between average human and student teacher accuracy in Table 1.

	Quadrants	**Columns**	**Both**
simple-features	0.91 (p=0.013)	0.59 (p=0.221)	0.59 (p=0.054)
salient-dinos	0.52 (p=0.286)	0.86 (p=0.027)	0.63 (p=0.037)

Table 1: Pearson correlations (with associated p-values) of average human student accuracy and representational alignment of the machine teacher across the various conditions.

Appendix E. Additional details on classroom simulations

E.1. Additional details on classroom pool construction

We explore two different pools of students and teachers: (i) unstructured pools spanning a range of representational alignments and error rates, and (ii) clustered sets of students and teachers. For the latter, we construct a generative model over student-teacher populations wherein we have a set of clusters, with a fixed number of students per cluster share similar representations. We sample one similar teacher from each cluster with some error rate (sampled from a uniform distribution over 0 to 0.5). We then deliberately downsample from the available teachers to simulate the case where some students are representationally distinct from the other students and available pool of teachers. Additional details are included in Appendix E.3. For each experiment, we sample 10 different teacher pools. We additionally compute the proportion of students who achieve "passing" marks (set to a moderately high threshold of 45% accurate, given chance guessing is 16.6% on our 6x6 grid). We also note that we focus here on row-based labels (a new f).

E.2. Utility curve over classrooms

The utility curves that we construct in Section 3 and D.2 were always constructed with respect to a single student (in a respective, "dyadic"[3]). However, in our classroom settings, we also corrupt the students' representations to simulate representational diversity. We

[3] Pairing two agents – one student and one teacher.

sample a new utility curve, wherein, for each teacher parameterization (same error rate parameterization as above, with representation corruptions now in increments of 0.1), we sample 10 different student corruptions ranging over 0 to 0.9 in increments of 0.1. We build this curve only for the column label type. We then bucketize the teacher error rate as well as the representation alignment such that we can index into the curve to extract an "expected average performance" for any student-teacher pair.

E.3. Constructing classroom pools

We construct two different classroom pools in Section 6: unstructured and structured. Here, we provide additional details on how we sampled students and teachers for each pool type.

Unstructured pool All students and teachers are sampled independently. We sample 1000 students and 30 teachers with corruption levels (pairwise swaps) sampled from a beta distribution ($\alpha = 1.5, \beta = 2.5$) to ensure that we have some students that are reasonably aligned. We sample teacher error rate uniformly over the range $0 - 0.5$.

Structured pool In the structured setting, we construct *clusters* of similar students and teachers. We prespecify a number of clusters M and number of students per cluster. Clusters are designed to span a range of levels of representation alignment over the "original" grid. We loop over possible representation alignments corruptions ranging from 0 to 1 in increments of $1/M$. For each cluster, we sample a "seed" student using that corruption level. We then sample students on top of this cluster with a representation corruption of 0.01 on top of the base student to ensure students share similar (but some variation) in their representation. For each cluster, we sample a teacher with error rate uniformly from $0 - 0.5$ and representation with a similar slight possible corruption (sampled uniformly from $0 - 0.01$) on top of the seed student, thereby ensuring that there *would* be a representationally similar teacher for each student in each cluster if provided. However, to simulate gaps in coverage of particular representation characterizations, we randomly drop some teachers from the pool.

E.4. Additional classroom matching results

We include additional results into classroom matching in Tables 2 and 3 and a relationship between group size and learning outcomes in Figure 5.

E.5. Generalization to the dino stimuli

We explore generalization of our utility curve constructed in the simple-features setting to our salient-dino stimuli. We repeat our two different pool types, which we depict in Tables 4 and 5, respectively. We find that our utility curves generalize nicely to different grid sizes and stimuli type, yielding student outcomes that on average appear to boost student accuracy particularly for the students in the top-performing group than baselines which do not account for representation misalignment (MOOC).

Method	Avg Acc	Bottom 10%	Top 10%	Pass Rate
Random	0.33 ± 0.01	0.21 ± 0.01	0.49 ± 0.01	0.12 ± 0.02
Min Err	0.37 ± 0.01	0.26 ± 0.01	0.53 ± 0.03	0.18 ± 0.04
Utility	**0.38 ± 0.01**	**0.27 ± 0.01**	**0.55 ± 0.03**	**0.20 ± 0.04**
Optimal	0.43 ± 0.01	0.32 ± 0.01	0.60 ± 0.02	0.32 ± 0.04

Table 2: Student learning outcomes (accuracy) from different classroom matching approaches in the structured pool setting. Higher is better for all metrics. ± indicates standard errors computed over 40 sampled pools and associated assignments. We compute the average student performance across all $N = 1000$ pooled students (paired with potentially $M = 30$ teachers), as well as accuracy over the bottom and top 10% of students in each matching, respectively. We additionally compute the proportion of students who achieve "passing" marks (set to a moderately high threshold of 45% accurate, given chance guessing is 16.6% on our 6x6 grid). Higher is better for all metrics. ± indicates standard errors computed over 10 sampled pools and associated assignments.

Method	Avg Acc	Bottom 10%	Top 10%	Pass Rate
Random	0.33 ± 0.00	0.17 ± 0.01	0.52 ± 0.01	0.09 ± 0.01
Min Err	**0.39 ± 0.01**	**0.25 ± 0.00**	0.57 ± 0.02	0.20 ± 0.02
Utility	**0.39 ± 0.01**	0.24 ± 0.00	**0.61 ± 0.02**	**0.23 ± 0.02**
Optimal	0.49 ± 0.00	0.36 ± 0.00	0.71 ± 0.02	0.54 ± 0.01

Table 3: Student learning outcomes (accuracy) from different classroom matching approaches in the unstructured pool setting. ± indicates standard errors computed over 40 sampled pools and associated assignments.

Method	Avg Acc	Bottom 10%	Top 10%	Pass Rate
Random	0.29 ± 0.00	0.16 ± 0.00	0.47 ± 0.01	0.04 ± 0.00
Min Err	0.35 ± 0.01	**0.22 ± 0.00**	0.55 ± 0.04	0.13 ± 0.03
Utility	**0.36 ± 0.01**	**0.22 ± 0.00**	**0.62 ± 0.04**	**0.16 ± 0.02**
Optimal	0.44 ± 0.01	0.31 ± 0.00	0.69 ± 0.03	0.33 ± 0.01

Table 4: Student learning outcomes (accuracy) from different classroom matching approaches in the unstructured pool setting for the dino stimuli. We again compute the student performance across all $N = 1000$ pooled students (paired with potentially 30 teachers). Error bars are again computed over 40 different sampled pools.

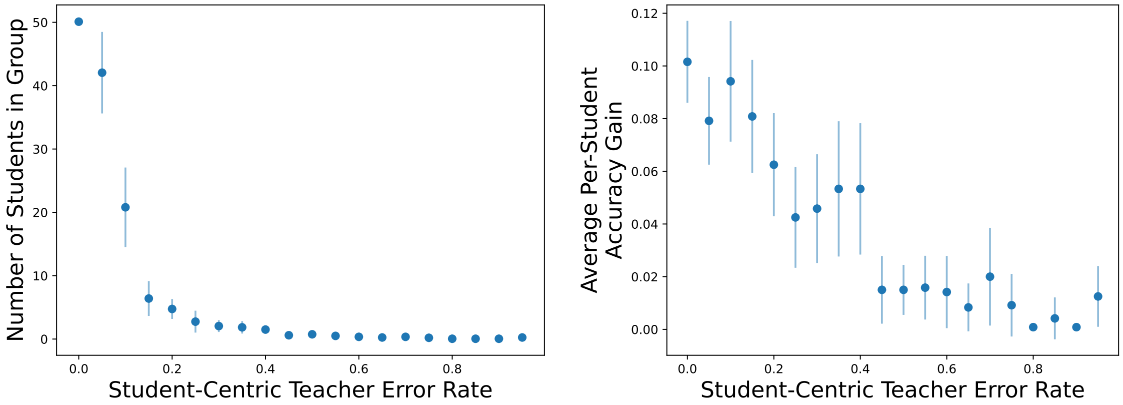


Figure 5: **(Left:)** Group sizes from greedily incorporating the lowest performing students into the classroom of a single student-centric teacher. **(Right:)** Average accuracy gains (out of 1.0) in performance for students grouped with the student-centered teacher, on top of what they would have achieved from a self-centered teacher. Error bars are standard errors over 20 seeds of student-centric teacher groupings for a sampled structured pool of students and teachers.

Method	Avg Acc	Bottom 10%	Top 10%	Pass Rate
Random	0.30 ± 0.01	0.18 ± 0.01	0.45 ± 0.02	0.06 ± 0.01
Min Err	0.34 ± 0.01	**0.23 ± 0.01**	0.51 ± 0.05	0.10 ± 0.03
Utility	**0.36 ± 0.01**	**0.23 ± 0.01**	**0.57 ± 0.05**	**0.15 ± 0.03**
Optimal	0.39 ± 0.01	0.28 ± 0.01	0.59 ± 0.04	0.18 ± 0.03

Table 5: Student learning outcomes (accuracy) from different classroom matching approaches in the structured (clustered) pool setting for the dino stimuli. We again have 10 representationally distinct clusters, each with 50 students, and sample 5 available teachers across the clusters.

E.6. Additional details on student-centric teacher

In contrast to our self-centered teacher, our student-centric teacher does not use its own representation to select examples to provide to the student. Instead, the student-centric teacher is endowed with an *inner optimization loop* over the students assigned to it, whereby the teacher loops T times over "simulated students" (which we call the "inner loop") and randomly selects one point per category (using the teacher's believed class – the teacher may not know the true categories) and measures the expected performance of each student if that set of examples were revealed. Note, the teacher computes the expected accuracy of each student using against its belief of the true categorization (which may be incorrect). The teacher then chooses the set of examples that attains the highest average accuracy over students. Here, we set T to 100; exploring the impact of varied T is a sensible next step. Exploration of alternate optimization functions, e.g., optimizing over the minimum attained performance over the students in the teacher's classroom rather than average classroom performance, as well as exploring different kinds of simulated students (here, we assume the teacher's have the right model of each student) are also ripe ground for future work.

We explore the effect of student-centric teachers by appending a second stage to our matching procedure. After matching using our utility curve (as noted above), we greedily attempt to pair the lowest performing students with a student-centric teacher who chooses points by optimizing for the students in their pool (i.e., taking the students' representations into account). We continue incorporating the next lowest-performing students into the student-centric teacher's classroom until a student's attained accuracy with the original pairing is not improved by the student-centric teacher. We apply our procedure to the clustered pool structure noted above and find that it is beneficial to continue adding students up to a point: if the teacher is an expert (zero error rate), we can add all students from one cluster before we see detrimental performance across the pool of students assigned to said teacher. As the student-centric teacher's error rate increases, fewer students can be pooled before performance dropoff (see Appendix Figure 5).

These results indicate the student-centric teachers can cover students who are representationally distinct and help boost their learning outcomes. However, classroom size matters, corroborating prior works in machine and human teaching (Frank, 2014; Yeo et al., 2019; Ma et al., 2018; Zhu et al., 2018). In the next section of the Appendix, we conduct a deeper dive into the relationship between classroom size and student outcomes in our setting when student-centric teachers are available. Herein, we see that teachers who may try to overalign to all students at once in a large classroom induce poorer outcomes for the classroom writ large.

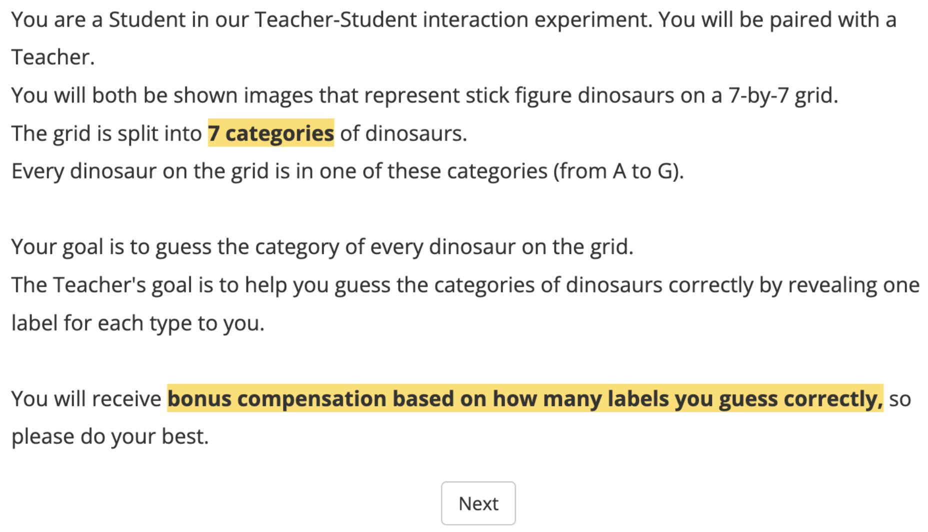
Appendix F. Additional human experiment details

F.1. Participant recruitment and compensation

Participants were recruited from Prolific and were paid \$12/hr plus a 10% bonus if they responded reasonably (i.e., did not select labels randomly or choose the same label for all stimuli). The research did not contain risks to participants, and they were able to opt out at any time. The institution of the principal investigator obtained IRB approval for this experiment, and participants gave informed consent under this protocol.

F.2. Task instructions

We include the full set of instructions provided to participants in Figure 6 and sample interfaces in Figure 7.



You are a Student in our Teacher-Student interaction experiment. You will be paired with a Teacher.
You will both be shown images that represent stick figure dinosaurs on a 7-by-7 grid.
The grid is split into **7 categories** of dinosaurs.
Every dinosaur on the grid is in one of these categories (from A to G).

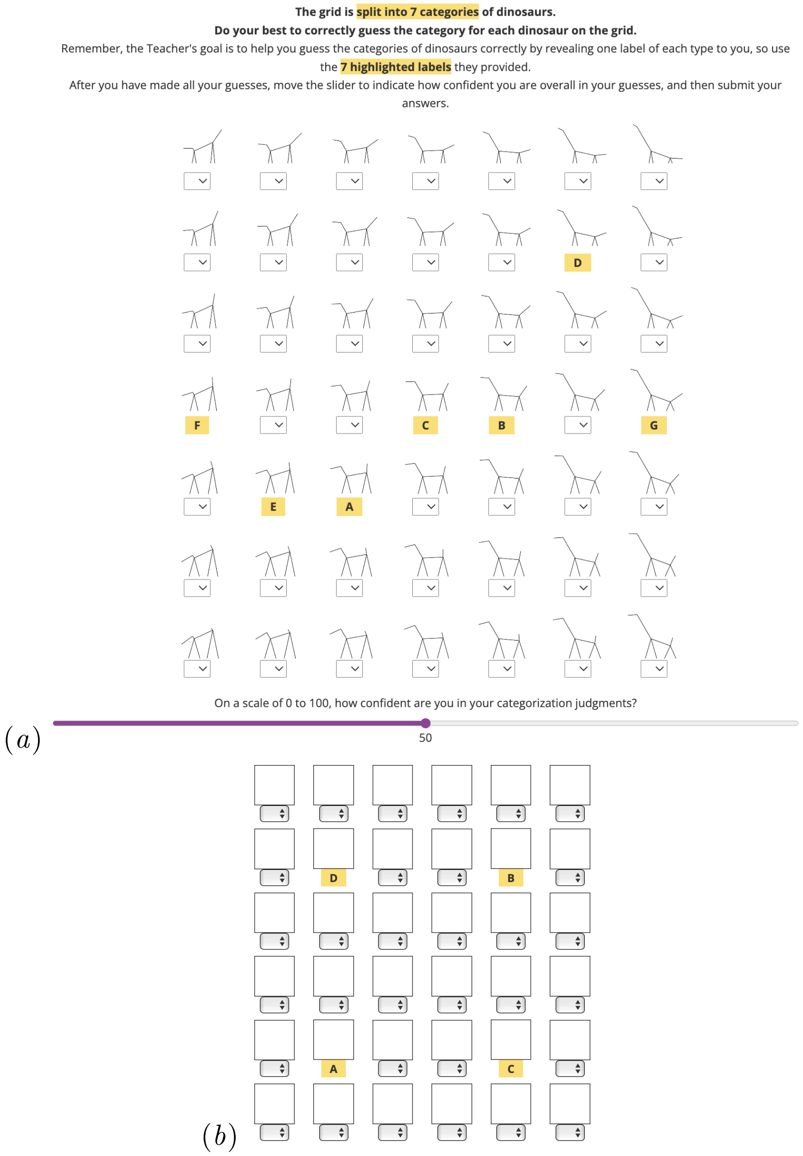
Your goal is to guess the category of every dinosaur on the grid.
The Teacher's goal is to help you guess the categories of dinosaurs correctly by revealing one label for each type to you.

You will receive **bonus compensation based on how many labels you guess correctly,** so please do your best.

Figure 6: Experiment instructions displayed to all participants, introduced paragraph by paragraph. The only changes to instructions were to modify the type of stimuli ("empty cells", "images that represent stick figure dinosaurs"), size of the grid (6×6, 7×7), the number and names of categories (4; A-D or 6/7; A to F/G).

F.3. Further analyses

Simulating teacher error in human experiments All human experiments were run with machine teachers set to zero error, as collecting all combinations of teacher error and representational alignment would be prohibitively expensive. Instead, we simulate the effect of teacher error in a post-hoc analysis by corrupting the true underlying labels in the same way we corrupted the teacher labels for the simulation experiments (i.e., error rate corresponds to the probability with which we flip each true label to be a different label). Human student accuracy was then recomputed against these corrupted true labels. The original human student results with no simulated teacher error are reported in Figure 8.



The grid is split into 7 categories of dinosaurs.

Do your best to correctly guess the category for each dinosaur on the grid.

Remember, the Teacher's goal is to help you guess the categories of dinosaurs correctly by revealing one label of each type to you, so use the **7 highlighted labels** they provided.

After you have made all your guesses, move the slider to indicate how confident you are overall in your guesses, and then submit your answers.

On a scale of 0 to 100, how confident are you in your categorization judgments?

50

Figure 7: Above are two example views of the experiment. All participants, after viewing the instructions in Figure 6 were taken to a page that contained a grid and the labeled stimuli. They were asked to categorize stimuli via a dropdown menu selection. Finally, they rated their confidence using a scale below the stimulus grid. **Left:** *salient-dinos*, 7 ("col") categories, medium-alignment teacher. **Right:** *simple-features*, 4 ("quad") categories, high-alignment teacher.

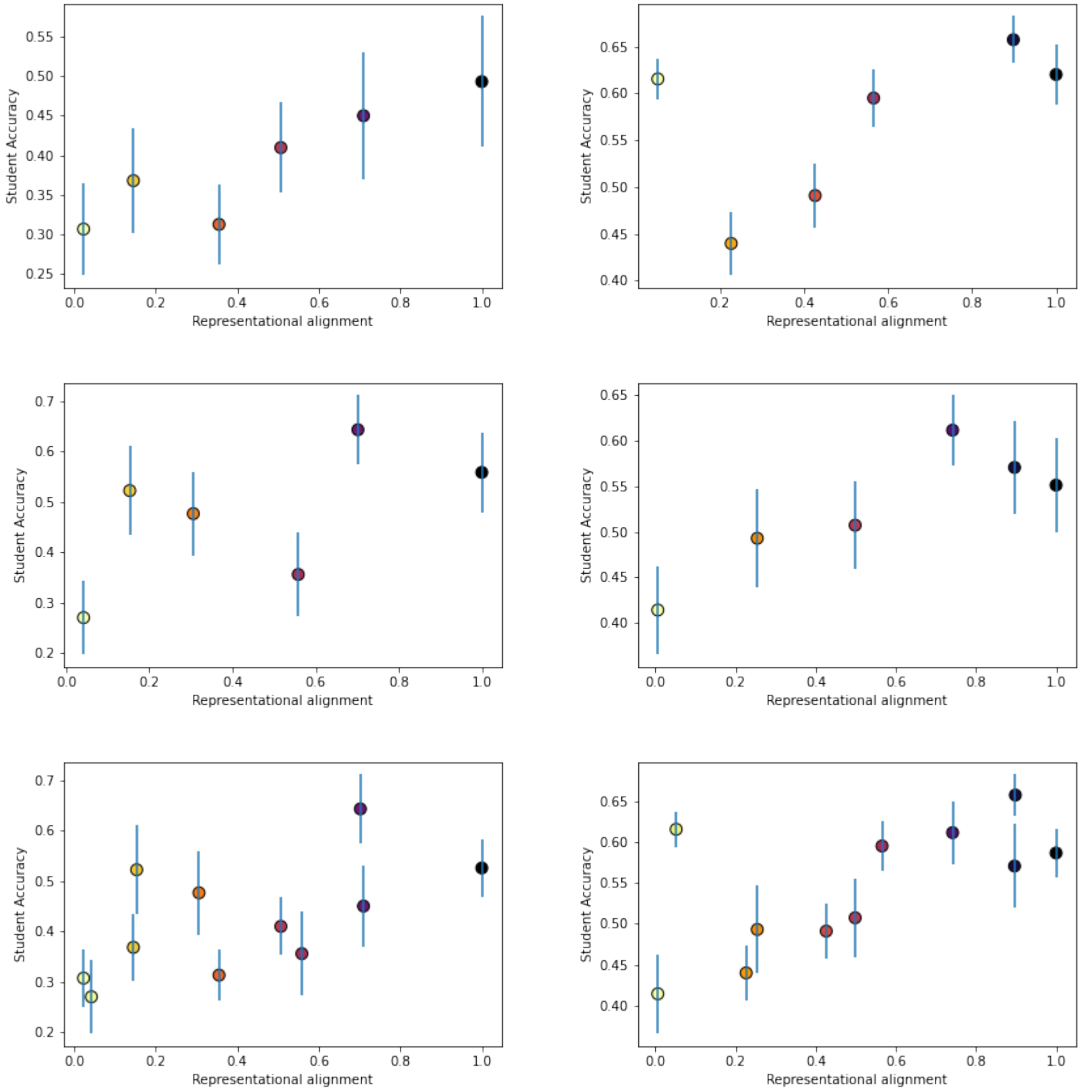


Figure 8: Average human student classification accuracy at various levels of representational alignment. Error bars correspond to one standard error. **(Left:)** Results from simple-features setting. **(Right:)** Results from salient-dinos setting. **(Top:)** One class per quadrant. **(Middle:)** One class per column (6 for simple-features, 7 for salient-dinos). **(Bottom:)** Combined results.