CleverBirds: A Multiple-Choice Benchmark for Fine-grained Human Knowledge Tracing

Leonie Bossemeyer[1] **Samuel Heinrich**[2] **Grant Van Horn**[3] **Oisin Mac Aodha**[1]
[1]University of Edinburgh [2]Cornell University [3]UMass Amherst

Abstract

Mastering fine-grained visual recognition, essential in many expert domains, can require that specialists undergo years of dedicated training. Modeling the progression of such expertize in humans remains challenging, and accurately inferring a human learner's knowledge state is a key step toward understanding visual learning. We introduce CleverBirds, a large-scale knowledge tracing benchmark for fine-grained bird species recognition. Collected by the citizen-science platform eBird, it offers insight into how individuals acquire expertize in complex fine-grained classification. More than 40,000 participants have engaged in the quiz, answering over 17 million multiple-choice questions spanning over 10,000 bird species, with long-range learning patterns across an average of 400 questions per participant. We release this dataset to support the development and evaluation of new methods for visual knowledge tracing. We show that tracking learners' knowledge is challenging, especially across participant subgroups and question types, with different forms of contextual information offering varying degrees of predictive benefit. CleverBirds is among the largest benchmark of its kind, offering a substantially higher number of learnable concepts. With it, we hope to enable new avenues for studying the development of visual expertize over time and across individuals.

1 Introduction

Dedicated practice and expert instruction from knowledgeable teachers are essential ingredients for students tasked with mastering new subjects and concepts. However, providing access to high quality, yet affordable, instruction at scale is time consuming and expensive [8]. As a result, researchers have looked towards alternative, computer-assisted [61], tools in an attempt to overcome these hurdles.

At the heart of an effective automated tutoring system is a computational model of the human learner. The goal of these models is to observe the learner as they engage with the teaching material at hand, represent the learners' knowledge state, and estimate any potential knowledge gaps they may have. This task, also known as knowledge tracing (KT), has a long history in the literature [3]. Early solutions modeled human mastery of the material being learned via latent variable probabilistic models [10, 11]. More recent approaches have advocated for the use of deep learning-based solutions [60] which, while effective at capturing more complex relationships, can require large quantities of data to train. However, current datasets for quantifying the performance of different KT methods are typically concentrated around a small number of subjects such as mathematics [12, 29, 50, 74], programming [57], and language learning [14, 40].

In this work, we attempt to address a gap in the existing available benchmark datasets for KT. This is motivated by the fact that there are a large number of domains where learners wish to learn visual identification skills, e.g., in medicine, art, and biology, to name a few. Many of the tasks in these domains can be posed as classification problems, where the human learner attempts to learn the decision boundaries between different concepts (i.e., different semantic classes). One such domain is animal species classification. For example, there are now a number of online platforms where

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

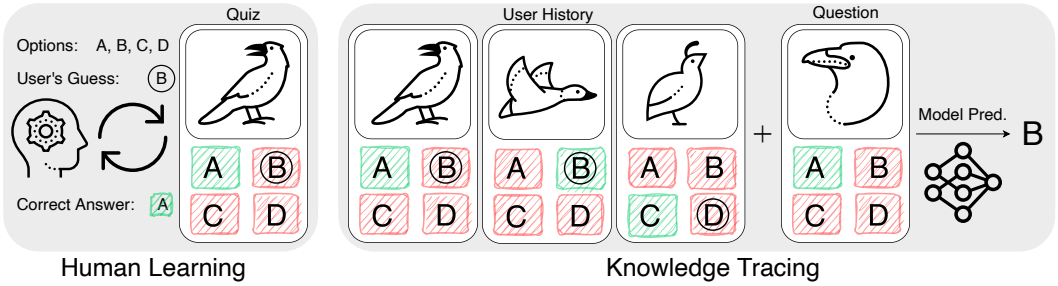

Figure 1: **(Left) Human Learning**. Participants learn from the quiz questions contained in CleverBirds through repeated interactions. For each question, participants are presented with an image of a bird species and a list of possible species names (here {'A', 'B', 'C', 'D'}), which may include the correct answer. After making a guess, they receive feedback in the form of the correct answer (here 'A'). This process is repeated for multiple questions. **(Right) Knowledge Tracing**. We illustrate the prediction task, in which a model is given a participant's interaction history together with the current question's image, options, and correct answer, and is tasked with predicting the participant's guess.

members of the public report sightings and locations of different species from all around the world, which in turn is providing valuable data for science [9]. The challenge for the participants in these projects is that the number of visual concepts (i.e., species) can be very large. For example, even if only restricted to the case of birds, there are over 11,000 different species worldwide. Compounding this difficulty is the fact that discriminating between certain species can require very fine-grained knowledge [75] as some species can look very similar to others.

To advance the development of KT methods in the context of fine-grained classification tasks we introduce the CleverBirds dataset. CleverBirds embodies a challenging real-world classification task and contains a large number of interactions generated by human participants who are attempting to learn how to identify different bird species from images. The core task is depicted in Fig. 1 where a learner is presented with a sequence of multiple-choice questions and the aim is for them to correctly identify the bird species depicted in the images shown. Example questions from the dataset are shown in Fig. 2.

We make the following contributions: (1) We introduce CleverBirds, a new large-scale benchmark for visual knowledge tracing. Our dataset contains over 10,000 visual unique concepts and more than 17 million total interactions from over 40,000 unique participants, with half of the participants having answered over 100 questions each. It provides a new benchmark for obtaining insights into human learning in the context of fine-grained visual classification. (2) We quantitatively evaluate a range of computational approaches on CleverBirds and demonstrate that it is a challenging benchmark, not only to human participants, but also to the computational methods tested. We evaluate these baseline methods under varying levels of input context, and show how different types of information can impact predictive accuracy. Links to the dataset and code are available at `https://cleverbirds-benchmark.github.io`.

2 Related Work

Knowledge Tracing (KT). KT methods aim to model student knowledge acquisition over time such that they can predict how a given student will perform on future interactions [3]. Effective models of human learning have a wide array of applications in the context of intelligent tutoring systems [61] and machine teaching [82]. Traditional KT approaches are based on probabilistic models of student mastery [15], but multiple extensions have been proposed to address some of the simplifying assumptions that were commonly made by earlier models, e.g., by incorporating individual-specific learnable parameters [78], by estimating concept difficulty [58], or by modeling more complex dependencies between concepts [39].

More recently, there has been a growth in the number of deep learning-based KT approaches proposed [65]. Various architectures have been explored such as recurrent networks [52, 60, 77], graph neural networks [53, 70, 76], attention-based models [32, 56, 57], memory augmented models [1, 81], hierarchical approaches [47, 73], and explainable methods [6]. There have also been attempts to

Figure 2: Three examples of the types of quiz questions found in our CleverBirds dataset. In each case, there are four options representing different species and an additional "None of the above option". The correct answer is indicated in green. Any of five options are valid answers and the set of candidate species provided in the option set are different for each question.

utilize contextual knowledge extracted from large language models to better encode interactions between concepts and questions [30, 45]. Deep KT methods can capture complex interactions and longer range temporal dependencies, but at the cost of requiring larger training datasets [31].

Datasets of Human Learning. There are a large number of benchmark datasets that have been utilized to quantify the performance of different KT methods. These datasets are typically distinguished in terms of the number of human learners (i.e., participants), the number of knowledge components/concepts (e.g., 'subtraction' could be a concept in the context of mathematics), the number of unique questions, and the total number of interactions (i.e., each student may not attempt all questions). Synthetic datasets have the advantage of enabling controlled testing as the underlying generative process is known [60]. However, evaluation in simulation is not a substitute for performing experiments using data from real human participants. The most common source of data comes from mathematical education, e.g., [29, 42, 66, 67, 74]. The largest of these datasets consist of millions of interactions with thousands of students and can also contain additional auxiliary information [50]. However, existing datasets are typically limited in the number of overall concepts contained within them. Beyond mathematics, other popular domains include linguistics [14], programming [30, 40], and general education games [40].

Most relevant to our benchmark are the small number of datasets targeting image classification tasks. While not precisely image data, [23] perform experiments on a dataset containing spectrograms derived from gravitational wave observations from the Gravity Spy citizen science project [80]. The goal for participants is to classify each spectrogram into one of a discrete set of classes representing 21 different types of 'glitches'. This dataset does not explicitly target a learning setting, i.e., it is not necessarily the case that the participants get better over time. The authors of [43] introduce three datasets for evaluating human learning of fine-grained visual concepts. Their datasets contain images of five species of butterflies, three classes of conditions of human retinas, and three classes of synthetic 'greebles'. Each dataset only contains 6,750 interactions which were obtained using participants from an online crowd working platform. For each of the three datasets, there are less than one thousand total images. We summarize the statistics of current KT datasets in Table A3.

In contrast to existing image-based datasets, our CleverBirds benchmark contains a much larger number of possible concepts that can be learned, i.e., 10,779 different species of birds from all around the world. In addition, the participant pool spans a range of expertize levels and is sourced from engaged individuals that volunteered to participate based on their interest in the problem domain.

3 CleverBirds Dataset

Here we describe our CleverBirds dataset. We outline the original collection protocol, the steps we undertook to refine it, and we describe the high-level statistics of the dataset.

Quiz Composition. The CleverBirds dataset is sourced from an online bird species identification quiz [16] created by the citizen science project eBird [68]. In the quiz, users (i.e., participants) are shown an image and asked to guess which bird species, from a list of options, is present in the image shown (see Fig. 2 for examples). It was first published online in March 2018 to encourage users to provide image quality ratings for new image uploads. The images contained within the quiz are sourced from citizen scientists who upload images, along with a proposed species label indicating the species present in the image, to the Macaulay Library [19]. This proposed species label is cross-referenced with a list of species expected to be found in the geographic location, and a volunteer expert reviewer is consulted for unlikely cases (i.e., an image identified as containing

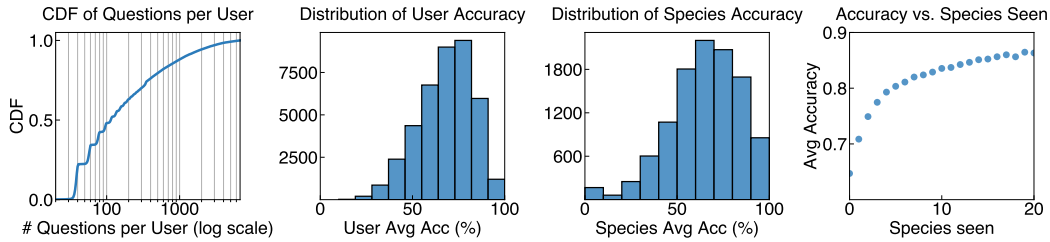

Figure 3: Left to right: Cumulative distribution of quizzes attempted per user on a log scale, distribution of users' average accuracies, distribution of species-wise average user accuracies, and average user accuracy by number of prior exposures to a species.

a specific species that is not typically found in that area). Data labeling errors from these types of citizen science efforts can occur, but are low. For example, on iNaturalist, the error rate for bird species identification has been found to be 3.3% [36].

Quiz images are sampled from images that have been uploaded within the last 5 to 365 days and contain proportions of both images with no quality ratings, and images that have quality ratings of at least 2.4 out of 5 possible stars. This allows for unrated images to be quality-labeled through use of the quiz, while keeping enough high quality images in the quiz for the users to learn from. Users are asked to rate image quality by sharpness, visibility of the bird, size of the photo and watermarks, while allowing for flocks and birds in-hand [21]. Additionally, users are encouraged to skip questions that are unanswerable, for example because multiple bird species are visible within one image. To further increase difficulty, the candidate answer options are selected to be taxonomically similar to the correct species. Specifically, options are drawn from a sliding window over the taxonomic list centered around the true species present in the image shown.

Graphical Interface. Users initiate a quiz by selecting parameters such as a location, time of year, and species prevalence, which are used to generate quiz questions. For example, if a user selected Edinburgh, Scotland, May 15th, and 'likely' species, they would only be presented with questions featuring common birds that would be expected to be found in that region at that time of year. Note, users have the option to select an audio quiz, whereby audio recordings are played instead of them being shown images. However, we only use the image quiz data as it is much more prevalent. Users are then guided through a set of 20 multiple-choice questions. For each question, an image of a bird is shown along with five answer choices: four species names and a "None of the above" option. Users can optionally skip any question. After submitting an answer, the correct species is revealed, and they are asked to rate the image quality on a scale of 1 to 5.

Data Filtering. CleverBirds is based on all quizzes completed online from March 14th 2018 to October 8th, 2024. To prevent overfitting on specific users, we split the dataset by user ID into training, validation, and test sets. Aiming for a 70/15/15 split, users are assigned to the training set until it comprises 70% of interactions, then we continue to the validation set and finally the test set. This results in 28,100 users in training, 6,021 in validation, and 6,023 in test, corresponding to 70.6%, 14.6%, and 14.8% of interactions, respectively. To respect image licenses, we do not provide the original images shown to users, but instead provide embeddings for each image using DINOv2 [55] with a ViT-B/14 backbone [28], as well as a ResNet50 [34] pretrained on ImageNet [25]. For DINOv2, we average-pooled the final layer's patch tokens after LayerNorm, excluding special tokens. For ResNet-50, we use the output of the global average pooling layer before the classifier.

To encode historical interactions, we construct a unique mapping for all species labels that appear in the dataset either as correct answers, or candidate options, and use it to encode the user's histories. In case of "None of the above" (NOTA) selections, the user's answer is encoded with a special token.

Dataset Characteristics. CleverBirds captures the learning dynamics of a diverse user population on a challenging real-world fine-grained classification task. It contains 17,859,392 user interactions, of which 98% involve unique image-species-choice combinations, and 26% contain unique species-choice pairs. 83% of images are never seen more than once, resulting in 14,753,114 distinct image features using the ImageNet ResNet50. For DINOv2, we provide 14,747,840 features, the discrepancy

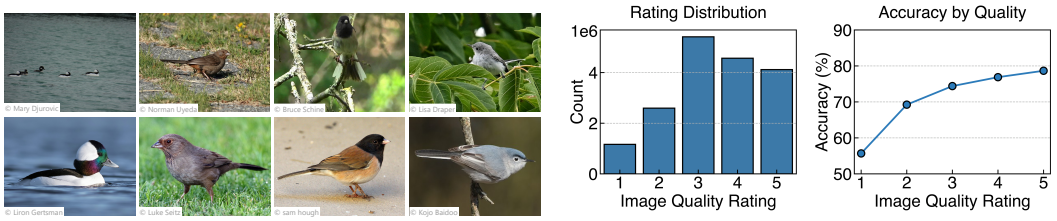

Figure 4: (Left) Here we compare lower quality quiz images (upper row) to high quality ones obtained from eBird species' pages (bottom row). Quiz questions may contain images that show birds from a distance, partially obscured, or uncommon angles. Species from left to right: Bufflehead, California Towhee, Dark-eyed Junco, and Blue-gray Gnatcatcher. (Right) Here we show the average accuracy of users for each possible quality rating. We observe that on average that higher quality images are easier for users.

(5,274) arising from the fact that DINOv2 was extracted later and images get deleted by the image owners over time. In our baselines, we treat unavailable images as zero inputs.

With over 10,000 species, CleverBirds spans a broad taxonomic range and exhibits wide variation in user engagement. As shown in the first panel of Fig. 3, over 50% of the 40,000+ users answered 100+ questions, and over 10% exceed 1,000 questions, enabling the analysis of learning dynamics over time. Users encounter a mean of 138 distinct species, and 10% encounter more than 300. Panel 2 and 3 of Fig. 3 show that the distribution of accuracies across users and species is broad, and centered around 60-70%. User improvement is observable, with over half of users showing measurable gains over sliding windows of 20 questions (one full quiz), making the dataset valuable for studying skill acquisition and feedback-driven learning. Panel 4 of Fig. 3 also illustrates this trend, as within the first 20 times a user sees a particular species, their accuracy increases by 20% on average. While focused on image-based identification, CleverBirds also contains substantial geographic diversity, with quiz selections drawn from over 4,000 distinct locations (see Fig. A2 for a visualization), and temporal coverage spanning all weeks of the year.

The difficulty of the task is further illustrated in Fig. 5, which displays the top-5 most frequently confused species pairs for species with over 1,000 interactions. For each species pair, the differences are subtle and require a trained eye to perceive. For example, the Pin-tailed Snipe can be differentiated from the Common Snipe by the white trailing edge of the wing of the Common Snipe [20]. The Sharp-shinned Hawk can be differentiated from the Cooper's Hawk by its smaller head, more squared-off tail, and smaller feet [22]. Note also, that the images shown for reference in Fig. 5 are high quality example images of the species. In the quiz, users could be confronted with partially obscured or zoomed out images of the same bird, increasing difficulty. Fig. 4 shows example quiz images compared to high quality images of the same species, highlighting the inherent difficulty within the quiz. Nevertheless, users achieve an average accuracy of over 50% even on images rated as low quality (see Fig. 4 - right panel). The distribution of image quality is shown in Fig. 4 - center panel.

Privacy. To ensure user privacy we anonymize all user-related identifiers, as well as those of quizzes, questions, and image assets that users interact with. We further aggregate quiz locations using the H3 geospatial index [71] at resolution 3 [72], which averages $12,393$ km^2 per cell. All quiz participants are registered Cornell Lab account holders and have agreed to the Terms of Use [18]. The project has been reviewed and approved by the School of Informatics Ethics Committee (project number 954242).

4 Problem Setting

Our goal is to model the learning progress of human learners as they engage with a multi-choice image classification quiz. At time $t \in \{1, \ldots, T\}$, the learner is presented with an image I_t and an ordered list of K possible candidate answers, denoted by $\mathbf{c}_t$. $\mathbf{c}_t$ contains a set of $K-1$ randomly ordered possible candidate answers, and also includes an addition none of the above option NOTA, yielding $\mathbf{c}_t = (c_{t,1}, \ldots, c_{t,K-1}, \text{NOTA})$.

We represent an image I_t with a fixed vision encoder $\mathbf{x}_t = f(I_t) \in \mathbb{R}^d$, and condition models on $\mathbf{x}_t$. The learner observes the image I_t and, based on their internal state, selects a response $r_t \in \mathbf{c}_t$ to the

Figure 5: Top-5 most frequently confused species pairs for species with > 1,000 interactions. From top-to-bottom and left-to-right: American Crow vs Fish Crow, Pin-tailed Snipe vs Common Snipe, Redpoll (Hoary) vs Redpoll (Common), Ross's Goose vs Snow Goose, Sharp-shinned Hawk vs Cooper's Hawk, and Short-tailed Shearwater vs Sooty Shearwater. Images taken from eBird [17].

question. The learner then receives the true species label y_t as feedback, and proceeds to the next question in the quiz. This process forms a single interaction $h_t = (\mathbf{x}_t, \mathbf{c}_t, y_t, r_t)$.

The learner's response is governed by their unobservable internal state θ_t, which summarizes their accumulated knowledge and memory, in conjunction with the input image and candidate choices shown. We assume the learner's state is updated after every interaction, similar to [43], and model the learner's response process as

$$r_t = \arg\max_{c \in \mathbf{c}_t} P(c \mid \mathbf{x}_t, \mathbf{c}_t, \theta_t), \tag{1}$$

where r_t denotes the categorical species selected by the learner and $r_t^{\text{bin}} = \mathbb{I}[y_t = r_t] \in \{0, 1\}$ indicates whether the answer was correct. Here, $P(\cdot)$ represents the learner's true (but unknown) response distribution conditioned on their internal state.

Given $P(\cdot)$ is unobserved, we approximate it via a shared parametric model ϕ trained across learners. ϕ does not include learner-specific parameters, instead, learner-specific behavior emerges through conditioning on each individual's recent interaction history $\mathcal{H}_t = (h_\tau)_{\tau=\max(1,\, t-W)}^{t-1}$, along with the current question $(\mathbf{x}_t, \mathbf{c}_t, y_t)$. Here, $W \in \mathbb{N}$ determines the maximum number of historical quiz questions that are assumed to influence the learner at a given time. While learners may in practice gain experience from other sources (e.g., in the wild observations), we exclude such influences and assume a direct correspondence between a learner's state θ_t and their interaction history $\mathcal{H}_t$. For some models, the conditioning set is further augmented by including features such as the learner's chosen quiz location and time focus, or by removing information such as the image features.

The model ϕ, as an approximation of $P(\cdot)$, estimates a response probability distribution and predicts either the categorical outcome as

$$\hat{r}_t = \arg\max_{c \in \mathbf{c}_t} \phi(c \mid \mathbf{x}_t, \mathbf{c}_t, y_t, \mathcal{H}_t), \tag{2}$$

or binary outcome as $\hat{r}_t^{\text{bin}} = \mathbb{I}[y_t = \hat{r}_t]$. The model is supervised to estimate either the response likelihood $\phi(r_t \mid \mathbf{x}_t, \mathbf{c}_t, y_t, \mathcal{H}_t)$ or, in the binary formulation, the correctness likelihood $\phi(r_t^{\text{bin}} = 1 \mid \mathbf{x}_t, \mathbf{c}_t, y_t, \mathcal{H}_t)$.

Images. As in [43], we use a fixed vision encoder to represent images for modeling. Individual learners may weight or attend to different features within these embeddings. We assume that learners first construct an internal visual representation of the image, which is subsequently used for species categorization, in line with studies indicating categorical readout from learned visual representations [4, 37]. Although this abstraction does not perfectly capture the learner's true visual representation, we hypothesize that human participants and neural networks trained on the same task extract similar image concepts. Due to their high dimensionality and pretraining on real-world images, we expect these features to encode much of the information used by both novice and expert learners. Prior work has demonstrated that pretrained CNN features can be predictive of human similarity judgments [5]. We present results on the predictive accuracy of these image features for species classification in Table A9.

5 Methods

We quantitatively evaluate CleverBirds across a range of different models ϕ predicting user responses given different levels of context contained in the quiz question: correct species y_t, choice candidates $\mathbf{c}_t$, image features $\mathbf{x}_t$, interaction history $\mathcal{H}_t$, and focus indicators i_{loc} and i_{st}. This context can be categorized into three main domains: (1) *User Context*: Features that are directly associated to a particular user, such as transcripts from their interaction history, or their previous performance and preferences. This context is aggregated from the interaction history $\mathcal{H}_t$ and focus indicators i_{loc} and i_{st}. (2) *Species Context*: Species-level features that are aggregated from the training set, such as average species difficulty of the correct species y_t and choice candidates $\mathbf{c}_t$. (3) *Image Context*: Extracted image features, implicitly encoding image concepts such as quality and ambiguity. Models relying on user, species, and image context are denoted with `U`, `S`, and `Img` respectively. Combinations are denoted with `U+S` or `U+S+Img`, indicating user and species, or user, species, and image context, respectively.

Evaluation. We evaluate the methods on a dataset of held out user IDs. The evaluation metrics for the binary task are binary macro accuracy which is the macro averaged accuracy over the correct outcome versus incorrect, and binary AP for correctly predicting user mistakes which is the average precision with the minority incorrect class treated as the "positive" class. For the multiple choice task, we report multiple choice accuracy which is accuracy over labels 1 to 5, where 5 denotes "None of the above", and multiple choice incorrect set accuracy which is the accuracy computed on the subset of questions that participants answered incorrectly, measuring how well the model recovers the correct label among distractors for questions participants fail to answer correctly.

Multiple-choice Classifiers. For multiple-choice response prediction, we evaluate transformer-based KT models, a confusion prior classifier, a simple MLP, and two heuristics. The first heuristic assumes an all-knowing learner that always selects the correct answer (*Always Correct*). The confusion prior classifier (*Conf Prior*) estimates the probability of each choice by masking the training-set confusion between the correct species and distractors, then re-normalizing to form a valid distribution over the choices presented to the user. We also add an additional confusion prior model which is constrained to predict only incorrect choices. This setup simulates a confusion prior focused exclusively on the subset of user questions answered incorrectly. As a lightweight neural baseline, we train a one-layer MLP that receives a learned 250-dimensional embedding of the correct species along with optional context (none, user-context, species-context, or both). The model outputs a probability distribution over all species, which is then masked to include only the five presented choices. We test user and species context (*MLP* `U+S`) and user, species, and image context (*MLP* `U+S+Img`). The image features are passed through an embedding layer converting them to the MLP input dimension. For both the confusion prior classifier and the one-layer MLP, NOTA is selected if the total probability assigned to species outside the available options exceeds the probability of every presented choice.

Binary Classifiers. For correct/incorrect prediction we fit simple probabilistic models (logistic regression (*LR*), XGBoost (*XG*), random forests (*RF*) with combinations of user and species context (`U`, `S`, and `U+S`). Additionally, we fit an average species classifier heuristic (*Avg Species*), which mirrors average species accuracy in the training set. A full description of models and their hyperparameters can be found in Appendix C.

Knowledge Tracing Models. We also evaluate several *knowledge tracing* baselines on our binary classification task. simpleKT [49] and Knowledge Query Networks (KQN) [44] serve as lightweight baselines. DKT [60], its regularized variant DKT+ [77], and the adversarial-training-based knowledge tracing (ATKT) [33], which adds adversarial training, model a learner's history using an LSTM hidden state. The self-attentive model for knowledge tracing (SAKT) [56] and attentive knowledge tracing (AKT) [32] selectively attend to the most relevant past interactions, with AKT using Rasch model [63] inspired regularization. Dynamic Key-Value Memory Networks for Knowledge Tracing (DKVMN) [81] externalizes knowledge into a key–value memory, enabling long-range tracking of per-concept knowledge.

We also evaluate two language modeling training paradigms for our multiple-choice task: sequence-to-sequence (LM-Seq2seq) generation [62] and multiple-choice classification (LM-MCC) [79]. For seq2seq, we fine-tune a pretrained T5-style [62] encoder-decoder transformer [69] to generate the correct species token based on the user's interaction history and the current question. For the classification model, we fine-tune a pretrained Bert-style [26] encoder-only transformer [38] to score the compatibility of each question-choice pair, conditioned on the same history.

For both models, we use a custom tokenizer which only includes one token per possible species, i.e., 11,142 tokens, along with special tokens for padding, segment separation, and token types. This avoids ambiguity when using natural language to describe species names, and ensures a compact token representation of our task, with each question occupying exactly 8 tokens.

For the seq2seq model, supervision is applied via a token-level cross-entropy loss on the generated species token, with NOTA aggregating all probabilities outside the visible choice candidates. For the classification model, binary labels are assigned to each candidate, with exactly one correct answer per question. For these transformer models, we set $W = 50$, other models use the complete history. These transformer models are evaluated using accuracy on the full dataset and subset of incorrect answers, as in case of the other multiple-choice models. Please see Appendix C for additional implementation details for these models.

Image Features. To showcase the use of image features for our prediction task, we evaluate an MLP with image features as input, with and without additional user and species context. The MLP using only image features (*MLP* `Img`), omitting user and species context, is similar to the static tracing model in [43]. Image features are encoded via an embedding layer to a dimension matching the number of unique species in the dataset. We report results for DINOv2 image features below, but full results, including using ResNet-50 image features, can be found in Table A1.

6 Results

We focus our discussion on high level takeaways from Fig. 6. Additional results (e.g., ResNet-50 image feature results, and additional KT baseline results) can be found in the appendix.

Feature engineered context result in strong baselines. The bottom row of Fig. 6 shows performance on the binary classification task, where the goal is to predict whether a participant will correctly answer a given question. Models trained specifically for this binary task outperform those trained for multiple choice and then subsequently binarized. The random forest model, leveraging both user and species context, performs best overall, achieving over 80% average precision in predicting participant errors. On the multiple-choice task, a one-layer MLP with user and species context matches the performance of significantly larger transformer models with a much larger capacity.

User context is more informative than species context. When comparing different context types, models receiving both user and species context (i.e., `U+S`) perform best overall, with models receiving only user context (i.e., `U`) close behind. Among the binary classifiers, those incorporating user context consistently achieve higher accuracy and average precision. While species context models still outperform simple heuristics on the binary task and full-dataset multiple-choice task, user-specific context appears necessary for strong predictions.

Multiple-choice trained classifiers are beaten by binary-trained classifiers on the binary task. On the binary task, multiple-choice classifiers are consistently outperformed by those trained directly on the binary objective. Despite the small capacity of our probabilistic models, they can achieve an AP and average accuracy of over 80%. By contrast, models originally trained to predict the exact multiple-choice response must allocate capacity not only to correctness but also to the structure of the incorrect responses. Future work could explore how these multiple-choice models could benefit from the outputs of binary classifiers, either as auxiliary signals or as gating mechanisms, or receive additional binary supervision.

Image features complement user and species context for incorrect choices. On the full dataset, the MLP with user and species context and DINOv2 image feature input shows a slightly higher performance compared to the other multiple-choice trained models (76%). On only incorrectly answered questions, however, it significantly outperforms the other parametric models and the random baseline with results around 25%. The same MLP without user and species context, *MLP* `Img`, achieves comparable results on the full dataset, but achieves only around 11% on the incorrect subset. This shows that image features help knowledge tracing for CleverBirds, especially if combined with the right context.

Predicting incorrect choices is challenging. Our trained models achieve approximately 70% accuracy on the multiple-choice task (see Fig. 6, top left panel). However, as observed in Fig. 3, participants perform substantially above random chance. The top right panel of Fig. 6 shows model accuracy on the subset of questions where participants select the *incorrect* answer. All trained models

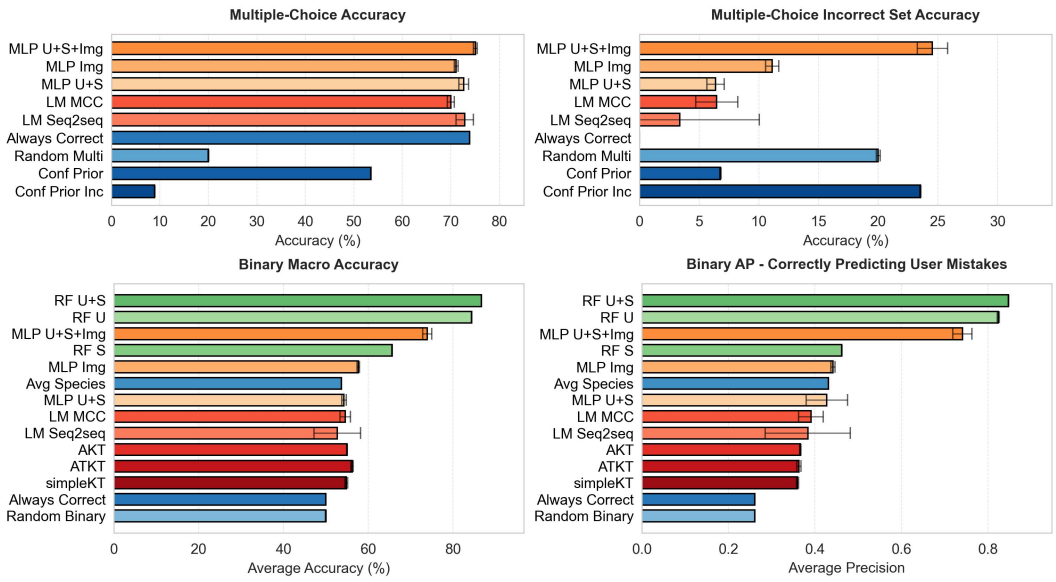

Figure 6: Performance on the multiple-choice and binary tasks. Top-left: accuracy on the full multiple-choice dataset. Top-right: accuracy on the subset of questions answered incorrectly. Bottom-left: macro-averaged accuracy on the binary task. Bottom-right: average precision (AP) for predicting user errors. Models are grouped by color into simple classifiers (*RF* `U`, *RF* `S`, *RF* `U+S`), MLPs (*MLP* `U+S+Img`, *MLP* `Img`), KT models (*LM MCC*, *LM Seq2seq*, *AKT*[32], *ATKT*[33], and *simpleKT* [49]), and simple heuristics (*Always Correct*, *Random binary*, *Random multiple-choice*, *Conf Prior*, *Conf Prior Inc*). *Img* uses DINOv2 features here, full results can be found in Appendix A.

achieve less than 25% accuracy on this subset, while the random classifier attains the expected 20% (i.e., a one-in-five chance), and the always-correct classifier scores 0% by design. The confusion prior, which assumes users are always incorrect (*Conf Prior Inc*), predictably underperforms on the full dataset (below 9% accuracy), but achieves over 23% accuracy on the incorrect subset. High performance on this subset would indicate that a model is able to approximate the participants' internal knowledge states. The marginal improvement of the best trained model over the incorrect-assumption confusion prior suggests that there is substantial room for improvement. Future work could explore the use of longer temporal contexts and better generalization across species, since success or failure on one species could provide information for others.

Species context can improve incorrect predictions. To isolate whether species context can guide incorrect predictions even without image features, we created an additional confusion prior classifier, *Conf Prior Inc*, and restricted it to never answer the correct species. As expected, this classifier performs poorly on the full dataset (accuracy of 10%). On the incorrect subset, it managed to outperform the random baseline, and obtain a similar accuracy as the MLP classifier with user, species, image context (around 24%). This shows that species context can help with predicting incorrect guesses.

Knowledge tracing baselines have little variance in performance. Table A2 summarizes the performance of the KT methods. As in [50], the tested KT methods all show similar performance, with average precision on predicting user's mistakes around 0.35 and average accuracy around 54%. This is surprising, given the diversity of model architectures. In terms of average accuracy, these models underperform compared to our simple classifiers incorporating user and species context. We note that most KT methods are designed for datasets with far fewer concepts than CleverBirds, and typically offer concept-level interpretability, which can limit their flexibility. As shown in Fig. 6, macro accuracy and AP (mistakes) are low for the KT methods AKT, ATKT, and simpleKT, despite high accuracy on the positive class. Similar results are shown for additional KT methods in Table A2. This suggests suboptimal performance on the negative class, which are instances where participants selected an incorrect species. One plausible explanation is that these models exploit a shortcut, focusing disproportionately on the positive class. We view this as evidence of CleverBirds' value in exposing limitations of existing approaches which motivates the development of more effective methods in future.

7 Limitations

While CleverBirds is the largest and most diverse dataset for visual knowledge tracing, it has some limitations. First, it focuses exclusively on bird species identification. We acknowledge this domain specificity limits immediate applicability to other areas such as medical imaging or object recognition. This is partially mitigated by the visual diversity of birds and the prevalence of hard-to-distinguish species pairs, making it representative of many fine-grained classification tasks. Moreover, the dataset's scale of nearly 11,000 fine-grained categories and over 40,000 participants with extensive learning trajectories provides unprecedented insights into human visual learning dynamics that, while domain-specific, may inform modeling approaches across fine-grained recognition tasks.

Second, the dataset reflects both location and selection biases. Participants are drawn from eBird [68], and participants on such platforms are known to be skewed towards the Global North [24]. See Appendix A.3 for additional details on geographic bias. The multiple choice format is a tradeoff that differs from open-ended identification, limiting knowledge tracing to a finite set of options and influencing ecological validity. This is partially mitigated through controlled design of the quiz, i.e., distractors are dynamically sampled from a sliding window centered on the true species, ensuring variability and substantial difficulty. Distractors are selected to be taxonomically similar, making the task challenging and realistic. A related constraint is the feedback, which only provides the correct species label to the user. Despite this simplicity, the signal is highly effective for learning in our fine-grained setting. Participant accuracy improves significantly with repeated species exposure (Fig. 3), showing that the correct label is a potent signal for correcting subtle classification errors. This finding is consistent with prior work showing humans can acquire visual expertise from label supervision alone [64].

Finally, although some label noise may be present in the quiz answers, we do not expect it to be substantial (Section 3), and the primary task is to predict user responses rather than the ground truth species labels. As with all KT applications, care must be taken when developing models from datasets such as ours, as inaccurate models could negatively bias future human learning.

8 Conclusion

We introduce CleverBirds, a new benchmark for evaluating models on the task of fine-grained visual knowledge tracing. CleverBirds contains rich interaction data from over 40,000 participants who are attempting to visually discriminate over 10,000 different bird species from all over the world. There are several properties that make our dataset well suited as a benchmark for this task, e.g., it contains a large number of interactions over time originating from participants of different skill levels, participants are not static as we observe their overall performance improves over time, and the concept space is large and challenging to master. We also demonstrate that CleverBirds poses a challenge for the computational methods tested, but the inclusion of additional context information improves performance. CleverBirds opens the door to future avenues related to modeling human knowledge acquisition in complex real world visual discrimination tasks, in addition to spurring development of methods for teaching such knowledge to learners.

Acknowledgments. We thank the eBird users for providing bird images, the Macaulay Library for curating and hosting this content, and the eBird participants who took part in the Photo and Sound Quiz. Media from the Cornell Lab of Ornithology | Macaulay Library was embedded as image features, published, used for baselines, and a subset was included for illustration in the paper. LB was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. OMA was in part supported by a Royal Society Research Grant.

References

[1] G. Abdelrahman and Q. Wang. Knowledge tracing with sequential key-value memory networks. In *Conference on research and development in information retrieval*, 2019.

[2] G. Abdelrahman, S. Abdelfattah, Q. Wang, and Y. Lin. DBE-KT22: A knowledge tracing dataset based on online student evaluation. *arXiv:2208.12651*, 2022.

[3] G. Abdelrahman, Q. Wang, and B. Nunes. Knowledge tracing: A survey. *ACM Computing Surveys*, 2023.

[4] F. G. Ashby and W. T. Maddox. Human category learning 2.0. *Annals of the New York Academy of Sciences*, 2011.

[5] M. Attarian, B. D. Roads, and M. C. Mozer. Transforming neural network visual representations to predict human judgments of similarity. In *Workshop on Shared Visual Representations in Human and Machine Intelligence at NeurIPS*, 2020.

[6] Y. Bai, J. Zhao, T. Wei, Q. Cai, and L. He. A survey of explainable knowledge tracing. *Applied Intelligence*, 2024.

[7] N. Bier, S. Moore, and M. Van Velsen. Instrumenting courseware and leveraging data with the open learning initiative (oli). In *International Learning Analytics and Knowledge Conference*, 2019.

[8] B. S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 1984.

[9] C. T. Callaghan, A. G. Poore, T. Mesaglio, A. T. Moles, S. Nakagawa, C. Roberts, J. J. Rowley, A. VergÉs, J. H. Wilshire, and W. K. Cornwell. Three frontiers for the future of biodiversity research using citizen science data. *BioScience*, 2021.

[10] H. Cen, K. R. Koedinger, and B. Junker. Is over practice necessary? - improving learning efficiency with the cognitive tutor through educational data mining. *Frontiers in artificial intelligence and applications*, 2007.

[11] H. Cen, K. Koedinger, and B. Junker. Comparing two irt models for conjunctive skills. In *ITS*, 2008.

[12] H.-S. Chang, H.-J. Hsu, and K.-T. Chen. Modeling exercise relationships in e-learning: A unified approach. In *Educational Data Mining*, 2015.

[13] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *International conference on knowledge discovery and data mining*, 2016.

[14] Y. Choi, Y. Lee, D. Shin, J. Cho, S. Park, S. Lee, J. Baek, C. Bae, B. Kim, and J. Heo. Ednet: A large-scale hierarchical dataset in education. In *AIED*, 2020.

[15] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 1994.

[16] Cornell University. eBird Photo and Sound Quiz Announcement, 2025. URL `https://ebird.org/news/ebird-photo-sound-quiz`. Accessed: 2025-05-01.

[17] Cornell University. eBird, 2025. URL `https://ebird.org/`. Accessed: 2025-05-01.

[18] Cornell University. Terms of Use, 2025. URL `https://www.birds.cornell.edu/home/terms-of-use/`. Accessed: 2025-05-11.

[19] Cornell University. Macaulay Library, 2025. URL `https://www.macaulaylibrary.org/`. Accessed: 2025-05-01.

[20] Cornell University. Pin-tailed Snipe - eBird, 2025. URL `https://ebird.org/species/pitsni`. Accessed: 2025-05-13.

[21] Cornell University. How to Rate Media in the Macaulay Library/eBird, 2025. URL https://support.ebird.org/en/support/solutions/articles/48001064392-rating-media#RatingPhotos. Accessed: 2025-05-13.

[22] Cornell University. Sharp-shinned Hawk - eBird, 2025. URL https://ebird.org/species/shshaw. Accessed: 2025-05-13.

[23] K. Crowston, C. Østerlund, T. K. Lee, C. Jackson, M. Harandi, S. Allen, S. Bahaadini, S. Coughlin, A. K. Katsaggelos, and S. L. Larson. Knowledge tracing to model learning in online citizen science projects. *Transactions on Learning Technologies*, 2019.

[24] B. H. Daru and J. Rodriguez. Mass production of unvouchered records fails to represent global biodiversity patterns. *Nature Ecology and Evolution*, 2023.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL Anthology*, 2019.

[27] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *EMNLP*, 2021.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[29] M. Feng, N. Heffernan, and K. Koedinger. Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction*, 2009.

[30] L. Fu, H. Guan, K. Du, J. Lin, W. Xia, W. Zhang, R. Tang, Y. Wang, and Y. Yu. Sinkt: A structure-aware inductive knowledge tracing model with large language model. In *CIKM*, 2024.

[31] T. Gervet, K. Koedinger, J. Schneider, and T. Mitchell. When is deep learning the best approach to knowledge tracing? *Journal of Educational Data Mining*, 2020.

[32] A. Ghosh, N. Heffernan, and A. S. Lan. Context-aware attentive knowledge tracing. In *International conference on knowledge discovery and data mining*, 2020.

[33] X. Guo, Z. Huang, J. Gao, M. Shang, M. Shu, and J. Sun. Enhancing knowledge tracing via adversarial training. In H. T. Shen, Y. Zhuang, J. R. Smith, Y. Yang, P. César, F. Metze, and B. Prabhakaran, editors, *Multimedia Conference*, 2021.

[34] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[35] L. Hu, Z. Dong, J. Chen, G. Wang, Z. Wang, Z. Zhao, and F. Wu. Ptadisc: A cross-course dataset supporting personalized learning in cold-start scenarios. In *NeurIPS Datasets and Benchmarks*, 2023.

[36] iNaturalist. Observation Accuracy Experiment, 2025. URL https://www.inaturalist.org/observation_accuracy_experiments/5?tab=research_grade_results. Accessed: 2025-05-01.

[37] X. Jiang, E. Bradley, R. A. Rini, T. Zeffiro, J. VanMeter, and M. Riesenhuber. Categorization training results in shape-and category-selective human neural plasticity. *Neuron*, 2007.

[38] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu. Tinybert: Distilling bert for natural language understanding. In *EMNLP*, 2020.

[39] T. Käser, S. Klingler, A. G. Schwing, and M. Gross. Dynamic bayesian networks for student modeling. *Transactions on Learning Technologies*, 2017.

[40] D. Kim, U. Lee, S. Lee, J. Bae, T. Ahn, J. Park, G. Lee, and H. Kim. ES-KT-24: A multimodal knowledge tracing benchmark dataset with educational game playing video and synthetic text generation. In *ITS*, 2025.

[41] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.

[42] K. R. Koedinger, R. S. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper. A data repository for the edm community: The pslc datashop. *Handbook of educational data mining*, 2010.

[43] N. Kondapaneni, P. Perona, and O. M. Aodha. Visual knowledge tracing. In *ECCV*, pages 415–431. Springer, 2022.

[44] J. Lee and D.-Y. Yeung. Knowledge query network for knowledge tracing: How knowledge interacts with skills. In *International conference on learning analytics and knowledge*, 2019.

[45] U. Lee, J. Bae, D. Kim, S. Lee, J. Park, T. Ahn, G. Lee, D. Stratton, and H. Kim. Language model can do knowledge tracing: Simple but effective method to integrate language model and knowledge tracing task. *arXiv:2406.02893*, 2024.

[46] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, et al. Deepseek-v3 technical report. *arXiv:2412.19437*, 2024.

[47] S. Liu, R. Zou, J. Sun, K. Zhang, L. Jiang, D. Zhou, and J. Yang. A hierarchical memory network for knowledge tracing. *Expert Systems with Applications*, 2021.

[48] Z. Liu, Q. Liu, J. Chen, S. Huang, J. Tang, and W. Luo. pykt: A python library to benchmark deep learning based knowledge tracing models. In *NeurIPS*, 2022.

[49] Z. Liu, Q. Liu, J. Chen, S. Huang, and W. Luo. simplekt: a simple but tough-to-beat baseline for knowledge tracing. In *ICLR*, 2023.

[50] Z. Liu, Q. Liu, T. Guo, J. Chen, S. Huang, X. Zhao, J. Tang, W. Luo, and J. Weng. Xes3g5m: A knowledge tracing benchmark dataset with auxiliary information. In *NeurIPS Datasets and Benchmarks*, 2023.

[51] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[52] K. Nagatani, Q. Zhang, M. Sato, Y.-Y. Chen, F. Chen, and T. Ohkuma. Augmenting knowledge tracing by considering forgetting behavior. In *World wide web conference*, 2019.

[53] H. Nakagawa, Y. Iwasawa, and Y. Matsuo. Graph-based knowledge tracing: modeling student proficiency using graph neural network. In *International conference on web intelligence*, 2019.

[54] OpenAI. Introducing GPT-4.1 in the API, Sep 2025. URL `https://openai.com/index/gpt-4-1`. [Online; accessed 29. Sep. 2025].

[55] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024.

[56] S. Pandey and G. Karypis. A self-attentive model for knowledge tracing. In *EDM*, 2019.

[57] S. Pandey and J. Srivastava. Rkt: relation-aware self-attention for knowledge tracing. In *CIKM*, 2020.

[58] Z. A. Pardos and N. T. Heffernan. Kt-idem: Introducing item difficulty to the knowledge tracing model. In *User Modeling, Adaption and Personalization*, 2011.

[59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *JMLR*, 2011.

[60] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein. Deep Knowledge Tracing. In *NeurIPS*, 2015.

[61] J. Psotka, L. D. Massey, and S. A. Mutter. *Intelligent tutoring systems: Lessons learned*. Psychology Press, 1988.

[62] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 2020.

[63] G. Rasch. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.

[64] A. Singla, I. Bogunovic, G. Bartók, A. Karbasi, and A. Krause. Near-optimally teaching the crowd to classify. In *ICML*, 2014.

[65] X. Song, J. Li, T. Cai, S. Yang, T. Yang, and C. Liu. A survey on deep learning based knowledge tracing. *Knowledge-Based Systems*, 2022.

[66] J. Stamper, A. Niculescu-Mizil, S. Ritter, G. J. Gordon, and K. R. Koedinger. Algebra i 2005-2006. challenge data set from kdd cup 2010 educational data mining challenge., 2010.

[67] J. Stamper, A. Niculescu-Mizil, S. Ritter, G. J. Gordon, and K. R. Koedinger. Algebra i 2006-2007. challenge data set from kdd cup 2010 educational data mining challenge., 2010.

[68] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological conservation*, 2009.

[69] Y. Tay, M. Dehghani, J. Rao, W. Fedus, S. Abnar, H. W. Chung, S. Narang, D. Yogatama, A. Vaswani, and D. Metzler. Scale efficiently: Insights from pretraining and finetuning transformers. In *ICLR*, 2022.

[70] S. Tong, Q. Liu, W. Huang, Z. Hunag, E. Chen, C. Liu, H. Ma, and S. Wang. Structure-based knowledge tracing: An influence propagation view. In *ICDM*, 2020.

[71] Uber Technologies, Inc. H3: Uber's Hexagonal Hierarchical Spatial Index, 2018. URL `https://www.uber.com/en-DE/blog/h3`. Accessed: 2025-05-11.

[72] Uber Technologies, Inc. Tables of Cell Statistics Across Resolutions | H3, 2025. URL `https://h3geo.org/docs/core-library/restable`. Accessed: 2025-05-11.

[73] T. Wang, F. Ma, and J. Gao. Deep hierarchical knowledge tracing. In *International Conference on Educational Data Mining*, 2019.

[74] Z. Wang, A. Lamb, E. Saveliev, P. Cameron, Y. Zaykov, J. M. Hernández-Lobato, R. E. Turner, R. G. Baraniuk, C. Barton, S. P. Jones, S. Woodhead, and C. Zhang. Diagnostic questions: The neurips 2020 education challenge. *arXiv:2007.12061*, 2020.

[75] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie. Fine-grained image analysis with deep learning: A survey. *PAMI*, 2021.

[76] Y. Yang, J. Shen, Y. Qu, Y. Liu, K. Wang, Y. Zhu, W. Zhang, and Y. Yu. Gikt: a graph-based interaction model for knowledge tracing. In *Machine learning and knowledge discovery in databases*, 2021.

[77] C.-K. Yeung and D.-Y. Yeung. Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Conference on learning at scale*, 2018.

[78] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon. Individualized bayesian knowledge tracing models. In *Artificial intelligence in education*, 2013.

[79] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*, 2018.

[80] M. Zevin, S. Coughlin, S. Bahaadini, E. Besler, N. Rohani, S. Allen, M. Cabero, K. Crowston, A. K. Katsaggelos, S. L. Larson, et al. Gravity spy: integrating advanced ligo detector characterization, machine learning, and citizen science. *Classical and quantum gravity*, 2017.

[81] J. Zhang, X. Shi, I. King, and D.-Y. Yeung. Dynamic key-value memory networks for knowledge tracing. In *International conference on World Wide Web*, 2017.

[82] X. Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *AAAI*, 2015.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. Our main contributions are the dataset described in Section 3 and benchmark results are described in Section 6.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, please see Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We include general notation for the task, but no theoretical proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the data processing pipeline and modeling pipeline in Section 3 and Section 5. Additionally, we upload our data processing after the anonymization step, and code, configurations and instructions to reproduce our experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, we provide the dataset and codebase, configurations and instructions.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Yes, please see Section 3 for data processing and splits and Section 5 for descriptions on model architectures and training.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report error bars and describe their construction in the supplementary material. For experiments involving transformer-based models and knowledge-tracing baselines, however, we only have a samples size of 3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, the compute configuration is described in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, we follow the NeurIPS Code of Ethics. The project is approved with application number 954242 by the University of Edinburgh. No personally identifiable data is published or used in the paper, and users of the bird quiz consent to their answers being saved by the platform. This is also described in Section 3. There are potentially harmful consequences if the dataset or models from this dataset are used for prediction on out-of-distribution user populations. They could be misrepresented and receive poor quality teaching assistance, impacting their learning.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the specia l circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Yes, societal impact is discussed in Section 7.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: There is no risk of misuse and private data being released with this anonymized dataset. There is no user location data being released, only user-chosen quiz locations, aggregated to a coarse scale of Hex 3 resolution 3, as described in Section 3.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Yes, we credited them in appropriate ways.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, we provide the appropriate documentation, and users on the eBird platform have consented to terms of user (described in Section 3).

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: Users of the eBird Quiz website [16] are accepting the terms of use, as described in Section 3. Screenshots of current instructions to users are provided in the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

Justification: Yes, ethics approvals are mentioned in Section 3, with number 954242.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were only used for editing purposes (language and code).

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

A Additional Results

A.1 Full Results

Table A1 and Table A2 report results on CleverBirds. Complementing Fig. 6, Table A1 includes models using ResNet-50 image features (*MLP* `Img ResNet-50` and *MLP* `U+S+Img Resnet-50`), logistic regression (*LR* `U`, *LR* `S`, *LR* `U+S`) and XGBoost (*XG* `U+S`). Table A2 reports additional results for the KT baselines AKT [32], simpleKT [49], DKT [60], DKT+ [77], DKVMN [81], KQN [44], ATKT [33], SAKT [56], SKVMN [1] on the binary task, evaluated using *AP (mistakes)*, *Macro Accuracy (%)*, *AUC* and *Accuracy (%)*. *AP (mistakes)* and *Macro Accuracy (%)* are directly comparable with columns 2 and 3 of Table A1, while *AUC* and *Accuracy (%)* are directly comparable with Table 2 of [50]. Both our results and those of [50] show little variability across model types.

Table A1: Baseline results on the multiple-choice and binary tasks. Columns (2) and (3) show performance on the binary classification task (correct/incorrect). Columns (4) and (5) show performance on the multiple-choice classification task. AP is average precision on predicting learners' mistakes, macro accuracy is average accuracy over the two classes (macro recall). 'Full' describes the full dataset, 'Incorrect' only incorrectly-answered questions. Models are: (1) Knowledge tracing models: sequence-to-sequence transformer model (*LM Seq2seq*), multiple-choice classification transformer model (*LM MCC*), *AKT* [32], *ATKT* [33] and *simpleKT* [49]; (2) MLPs using user context (U), species context (S), and image features from ResNet-50 (*Img ResNet-50*) and DINOv2 (*Img DINOv2*) features; (3) Random Forest (*RF*), Logistic Regression (*LR*) and XGBoost (*XG*); (4) Random binary and multiple-choice baselines, and heuristics using the average species performance (*Avg Species*), user-correct assumption (*Always Correct*), confusion prior (*Conf Prior*) and the confusion prior which assumes incorrect participant guesses (*Conf Prior Inc*). Error bars are 2-sigma, based on 3 training runs for KT baselines, and 10 training runs for all other models. Metrics are reported as averages, with 2-sigma standard deviation.

Task	**Binary**	**Binary**	**Multiple Choice**	**Multiple Choice**
Metric	**AP**	**Macro Accuracy (%)**	**Accuracy (%)**	**Accuracy (%)**
Dataset	**Full**	**Full**	**Full**	**Incorrect**
LM Seq2seq	0.384 ± 0.098	66.90 ± 21.60	72.90 ± 1.80	3.40 ± 6.60
LM MCC	0.391 ± 0.028	66.60 ± 18.20	70.00 ± 0.60	6.50 ± 1.80
AKT	0.366 ± 0.002	55.00 ± 0.00	–	–
ATKT	0.363 ± 0.004	56.20 ± 0.40	–	–
simpleKT	0.360 ± 0.002	54.80 ± 0.40	–	–
MLP U+S	0.428 ± 0.048	67.60 ± 19.20	72.70 ± 1.00	6.40 ± 0.80
MLP U+S+Img ResNet-50	0.768 ± 0.042	79.90 ± 9.60	$\mathbf{76.00 \pm 1.20}$	$\mathbf{24.50 \pm 2.40}$
MLP U+S+Img DINOv2	0.741 ± 0.022	79.90 ± 8.80	75.10 ± 0.40	$\mathbf{24.50 \pm 1.20}$
MLP Img ResNet-50	0.461 ± 0.004	68.60 ± 18.00	72.90 ± 0.20	9.00 ± 0.40
MLP Img DINOv2	0.442 ± 0.004	68.70 ± 15.80	71.10 ± 0.40	11.10 ± 0.60
RF U	0.824 ± 0.002	81.40 ± 4.20	–	–
RF S	0.462 ± 0.000	64.80 ± 1.20	–	–
RF U+S	$\mathbf{0.848 \pm 0.000}$	84.90 ± 2.60	–	–
LR U	0.809 ± 0.002	80.60 ± 5.20	–	–
LR S	0.433 ± 0.000	64.70 ± 1.00	–	–
LR U+S	0.824 ± 0.002	83.80 ± 3.40	–	–
XG U+S	0.845 ± 0.000	$\mathbf{85.40 \pm 4.00}$	–	–
Random Binary	0.261 ± 0.000	50.00 ± 0.00	–	–
Random Multi	0.261 ± 0.000	40.40 ± 13.60	20.00 ± 0.00	20.00 ± 0.20
Avg Species	0.432 ± 0.000	67.60 ± 19.80	–	–
Always Correct	0.261 ± 0.000	65.90 ± 22.80	73.90 ± 0.00	0.00 ± 0.00
Conf Prior	0.252 ± 0.000	55.50 ± 9.80	53.50 ± 0.00	6.80 ± 0.00
Conf Prior Inc	0.245 ± 0.000	34.70 ± 20.80	8.90 ± 0.00	23.50 ± 0.00

Table A2: Results of knowledge tracing models on the binary classification task. Metrics are *AP*, *macro accuracy (%)*, *AUC* and *accuracy (%)*. Averages are reported with 2-sigma standard deviation on 3 runs. The models receive user context and predict binary correct/incorrect outcomes, the AP and Macro Accuracy values can thus be compared to columns (2) and (3) of Table A1.

Model	**AP (mistakes)**	**Macro Accuracy (%)**	**AUC**	**Accuracy (%)**
AKT [32]	**0.366** $\pm$ 0.001	54.976 $\pm$ 0.068	0.632 $\pm$ 0.002	72.293 $\pm$ 0.115
simpleKT [49]	0.360 $\pm$ 0.002	54.803 $\pm$ 0.262	0.628 $\pm$ 0.001	72.034 $\pm$ 0.013
DKT [60]	0.334 $\pm$ 0.002	54.137 $\pm$ 0.113	0.600 $\pm$ 0.001	70.369 $\pm$ 0.118
DKT+ [77]	0.333 $\pm$ 0.000	53.771 $\pm$ 0.000	0.600 $\pm$ 0.000	70.620 $\pm$ 0.000
DKVMN [81]	0.349 $\pm$ 0.003	54.941 $\pm$ 0.137	0.613 $\pm$ 0.002	71.017 $\pm$ 0.337
KQN [44]	0.340 $\pm$ 0.000	54.438 $\pm$ 0.297	0.604 $\pm$ 0.004	70.694 $\pm$ 0.505
ATKT [33]	0.363 $\pm$ 0.004	**56.183** $\pm$ 0.248	**0.636** $\pm$ 0.004	70.977 $\pm$ 0.229
SAKT [56]	0.342 $\pm$ 0.002	54.468 $\pm$ 0.185	0.604 $\pm$ 0.002	71.103 $\pm$ 0.224
SKVMN [1]	0.350 $\pm$ 0.001	52.920 $\pm$ 0.229	0.608 $\pm$ 0.001	**73.138** $\pm$ 0.174

Table A3: Comparison of existing knowledge tracing datasets. Table extended from [40].

Dataset	Participants	Questions	Concepts	Interactions	Subject
Simulated-5 [60]	4,000	50	5	200,000	Synthetic
ASSISTments2009 [29]	4,217	26,688	123	346,860	Math
ASSISTments2012 [29]	46,674	179,999	265	6,123,270	Math
ASSISTments2015 [29]	19,917	100	-	708,631	Math
ASSISTments2017 [29]	1,709	3,162	102	942,816	Math
Statistics2011 [7]	333	1,224	-	194,947	Math
Junyi2015 [12]	247,606	722	41	25,925,922	Math
KDD2005 [66]	574	210,710	112	809,694	Math
KDD2006 [67]	1,146	207,856	493	3,679,199	Math
XES3G5M [50]	18,066	7,652	865	5,549,635	Math
Eedi Task 1 [74]	118,971	27,613	-	15,867,850	Math
Eedi Task 2 [74]	4,918	948	-	1,382,727	Math
ES-KT-24 [40]	15,032	182	28	7,783,466	Math and language
POJ [57]	22,916	2,750	-	996,240	Programming
PTADisc [35]	1,530,100	225,615	4,054	680M+	Programming
Programming [30]	2,756	726	82	193,284	Programming
EdNet [14]	1,677,583	52,676	962	372,366,720	Linguistics
DBF-KT22 [2]	1,361	212	98	167,222	Computer and information science
GravitySpy [23]	10,655	51,047	21	1,026,652	Spectrogram classification
VTK-Greebles [43]	150	1,200	3	6,750	Synthetic image classification
VTK-Eyes [43]	150	600	3	6,750	Retinal disease classification
VTK-Butterflies [43]	150	2,224	5	6,750	Butterfly species classification
CleverBirds (Ours)	40,144	17,859,392	10,779	17,859,392	Birds species classification

A.2 In-Context Learning

To evaluate whether large language models can solve the task in-context, we evaluated OpenAI's GPT-4.1 Nano [54] and DeepSeek's DeepSeek-V3-0324 [46] on the multiple-choice task, using a subset of 100,000 test set examples. Each model received a natural language task description, the participant's interaction history, and a prompt for classification. Table A5 shows the exact prompts used. We use [HISTORY] to denote the participant's interaction history, formatted for sequence-to-sequence prediction and rendered with actual species names in text rather than tokenized codes. The history sequence ends with the current question, encouraging the model to input the participant's answer.

The results (Table A4) broadly reflect those of the fine-tuned models, i.e., GPT-4.1 Nano achieves high accuracy on correctly answered examples but struggles on incorrect ones, whereas DeepSeek exhibits lower overall accuracy but outperforms on the subset of incorrect examples. Comparing this to the other models in Fig. 6, GPT-4.1 Nano matches LM MCC on the full set, but underperforms the parametric models on the subset of incorrectly-answered questions. Deepseek-V3-0324 underperforms the other language models on the full test set, but outperforms them on the incorrectly-answered questions. We note two key limitations: we only evaluate a random subset of 100,000 examples, and we use identical prompts for both models without model-specific optimization.

Table A4: In-context multiple-choice accuracy on full dataset and incorrect subset using pre-trained cloud models for a random subset of 100,000 examples.

Task	**Multiple Choice**	
Metric	**Accuracy**	
Dataset	**Full**	**Incorrect**
Count	100,000	26,115
deepseek-chat [46]	0.53	**0.13**
gpt-4.1-mini-2025-04-14 [54]	**0.71**	0.02

Table A5: Prompts used for in-context learning task. [HISTORY] is replaced with the concatenation of all historical questions and the current question.

Component	**Content**
System Prompt	You are analyzing a user's quiz-taking behavior. In each quiz question, the user is shown an image and must choose the correct species from multiple-choice options. There are always five choices available: Four eBird species options, listed in the 'Options' field (indices 0–3). A fifth option labeled 'None of the Above', which is always choice 4. If the correct species is not among the four listed options, then 'None of the Above' (index 4) is the correct choice. Note: The 'None of the Above' option is always available but not shown in the 'Options' list. [HISTORY]. Your prediction should consider both the user's interaction history and the difficulty of the current and previous questions. Use your knowledge of bird species when relevant. Only reply with the correct digit of the choice (0, 1, 2, 3, or 4). Your answer (single digit only):
Historical Questions	Question: Correct answer: Northern Mockingbird Options: Northern Mockingbird | Eastern Bluebird | Western Bluebird | Mountain Bluebird User's answer: Northern Mockingbird
Current Question	Question: Correct answer: Woodhouse's Scrub-Jay Options: Canada Jay | Pinyon Jay | Steller's Jay | Blue Jay User's answer:

A.3 Performance Inside and Outside the US

To further analyse potential geographic biases, we compare performance across quizzes that were selected by participants as either inside or outside the US. We compare both the performance of participants on the quiz, as well as the knowledge tracing capability of our trained models.

In Table A6 we observe that participants exhibit no substantial performance difference when answering quizzes based on locations inside the US or outside it. In contrast, Table A7 shows that our models show a small but statistically significant performance gap favoring quizzes where the quiz location is in the US.

Table A6: Comparison of participant's performance on quiz locations inside and outside the US. Error bars are 2-sigma.

Location	**Count**	**Mean**
Inside US	6,808,489	**0.75** $\pm$ 0.43
Outside US	11,050,903	0.74 $\pm$ 0.44

Table A7: Model performance by quiz location. RF is the random forst model with user and species context (*RF* `U+S`), MLP is the MLP with user, species and ResNet image features (*MLP* `U+S+Img ResNet`). Error bars are 2-sigma.

Location	Model	Binary - AP	Binary - Macro Acc	MC - Acc Full	MC - Acc Inc
Inside USA	RF	**0.86** (±0.00)	**85.34** (±0.01)	–	–
Outside USA	RF	0.84 (±0.00)	83.27 (±0.01)	–	–
Inside USA	MLP	**0.79** (±0.02)	**84.30** (±0.59)	**76.71** (±0.33)	**25.90** (±1.94)
Outside USA	MLP	0.71 (±0.02)	81.92 (±0.34)	73.41 (±0.28)	23.98 (±1.23)

A.4 Learning Dynamics

To gain insights into learning dynamics, we investigated model performance given different amounts of participant context. For this, we sorted the test set by context length and divided it into quintiles on a logarithmic scale. Table A8 shows the result of our best performing binary model RF `U+S`. This reveals a clear monotonic improvement in both AP and macro accuracy as the amount of user context increases, suggesting that the model effectively integrates participant context into its predictions. Notably, the incremental gains decrease, with the sharpest rise observed between Q1 and Q2 and the smallest between Q4 and Q5, indicating diminishing returns of added context.

Table A8: Binary task performance of the random forest with user and species context (RF `U+S`) on the test set, stratified by context length quintiles on a logarithmic scale. Error bars are 2-sigma.

Metric	Q1	Q2	Q3	Q4	Q5
Binary AP	0.62 (±0.0002)	0.77 (±0.0002)	0.88 (±0.0001)	0.93 (±0.0001)	**0.96** (±0.0001)
Binary Macro Accuracy	67.25 (±0.0247)	80.09 (±0.0173)	87.73 (±0.0103)	91.73 (±0.0075)	**93.56** (±0.0090)

A.5 Image Classification

To understand how informative our image features are, we evaluate multiple-choice species classification (i.e., not participant guesses) directly from image features by training an one-layer MLP to predict the true species. The setup matches the image-only MLP (*MLP* `Img`) runs, except that the model is not provided the ground-truth species and the target is the true species rather than the participant's guess. Results are presented in Table A9. DINOv2 features achieve 88.8 ± 0.2% accuracy, outperforming ResNet-50 by 13.9 percentage points (74.9 ± 0.4%) with low variability across runs, indicating a strong advantage for the DINOv2 features.

Table A9: Species classification on the Multiple Choice task using an MLP with ResNet-50 or DINOv2 features. Error bars are 2-sigma over 10 training runs.

Metric	ResNet-50	DINOv2
Multiple Choice Accuracy (%)	74.90 ± 0.40	**88.80 ± 0.20**

B Additional Dataset Details

In CleverBirds, the order of choices is random, as can be seen in Fig. A1. Participants' guesses, on the other hand, slightly disfavor the "None of the above option" option, and slightly favor the fourth option. Quiz location are chosen around the globe, as shown in Fig. A2, which shows quizzes' at H3 Hex 3 locations with interaction density. The distribution favors populated areas, and the global north. Fig. A3 shows four screenshots of the quiz website. First, the participant sees a general

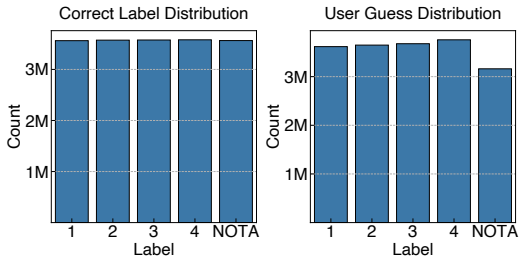

Figure A1: Distributions of true labels across quiz choices and participant responses.

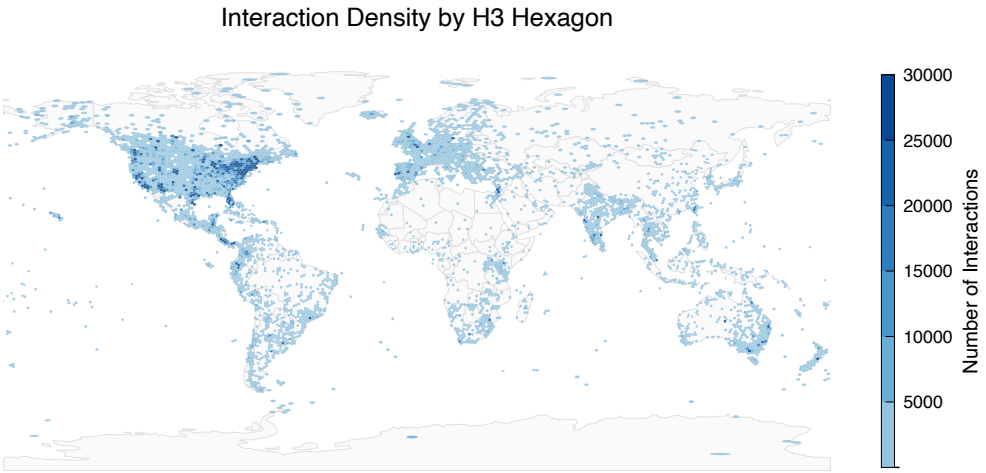

Figure A2: World map with Hex 3 polygonal bins representing quiz locations, where color intensity encodes the number of interactions per location cell.

introduction to the quiz. Next, participants configure the quiz by selecting the place, time, and media type. Each of the 20 quiz questions presents a bird image alongside five answer choices; after responding, participants receive feedback and may rate the image quality.

Existing Knowledge Tracing Datasets. In Table A3 we summarize existing knowledge tracing datasets. We can see that our CleverBirds dataset is significantly larger and more diverse in terms of the number of concepts contained withing compared to existing visual knowledge tracing datasets.

C Additional Implementation Details

Input Encoding Transformer Models For learner l, each historical interaction $h_s^l \in \mathcal{H}_t^l$ is encoded as:

$$h_s^l = (\mathsf{TOK}^C,\ y_s^l,\ \mathsf{TOK}^O,\ \mathbf{c}_s^l,\ \mathsf{TOK}^A,\ r_s^l), \quad \text{for } s \in [\max(0, t-W), t)], \tag{3}$$

where TOK^C, TOK^O, and TOK^A are defined as special tokens and are used to encode the type of information, specifically correct, options, and answers respectively, encoded in the next tokens. The model input consists of all interactions in the lookback window followed by the current question:

$$s_t^l = (h^l_{\max(0,t-W)},\ \mathsf{TOK}^S,\ h^l_{\max(0,t-W)+1},\ \mathsf{TOK}^S,\ \ldots,\ h^l_{t-1},\ q_t^l), \tag{4}$$

where TOK^S indicates the separation token. Here, the current question q_t^l includes the correct species, the available choices and a special prompt token, but crucially not the actual participant answer:

$$q_t^l = (\mathsf{TOK}^C,\ y_t^l,\ \mathsf{TOK}^O,\ \mathbf{c}_t^l,\ \mathsf{TOK}^U), \tag{5}$$

where TOK^U represents the participant's answer token type.

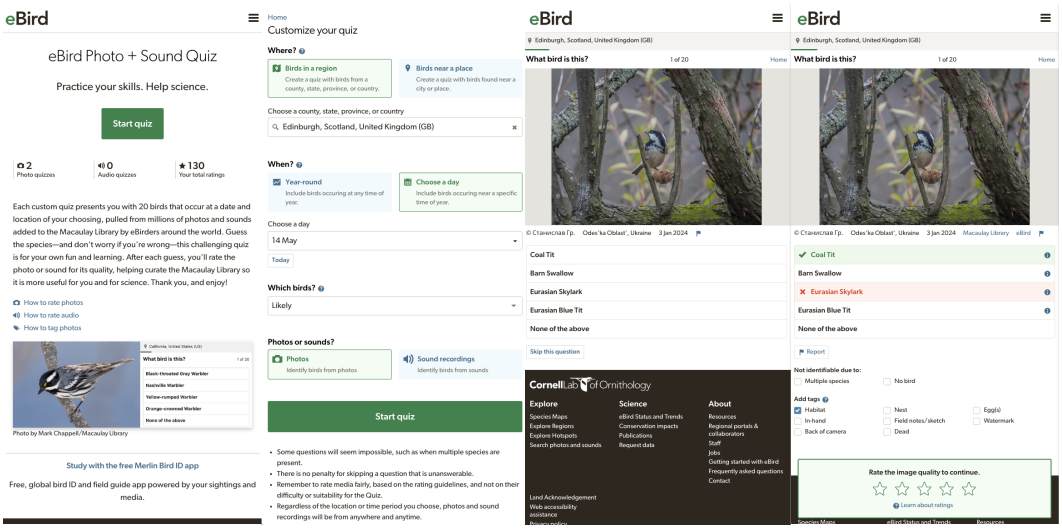

Figure A3: Screenshots of the participant interface from the eBird quiz [16] From left to right: (1) Quiz introduction page, presenting the task and linking to rating guidelines; (2) Customization page, where participants select quiz parameters such as location, time, species prevalence, and media type; (3) Question page, showing a bird image with five label choices and a skip option; (4) Feedback page, revealing the correct label and prompting a quality rating of the image.

In the classification model, each candidate answer $c^l_{t,j}$ is appended to the question input, resulting in $s^l_{t,j} = (q^l_t,\ c^l_{t,j})$, for $j = 1, \ldots, K$. A classification token is prepended to the full sequence before encoding.

Training Configuration. For the seq2seq model, we use Google's T5-Efficient-TINY-NL32 [69], pretrained on the English C4 Corpus [27]. For the MCC approach, we use TinyBERT [38], which is distilled from pretrained and fine-tuned BERT [26] teacher. *LM Seq2seq* and LM MCC are consequently on a server equipped with 2 AMD EPYC 7763 64-Core Processor 1.8GHz (128 cores in total, 32 used for training), 1TiB RAM and 4 NVIDIA A100-SXM-80GB GPUs, with bf16 precision.

We train each language model for 10 hours, using AdamW [51] with a learning rate of 5×10^{-5}, weight decay of 0.01, maximum gradient norm of 1, and 4 gradient accumulation steps. Seq2seq models use a batch size of 512, while MCC models use 256. We evaluate every 200 steps on a validation subset of 300,000 examples, and select the best epoch based on validation performance.

The MLP models (*MLP* `Img`, *MLP* `U+S+Img` and the image classification MLP models) are trained with hidden size 250, batch size 65,792, learning rate 0.001, and Adam [41] for 3 epochs. They train on GPU in 2 hours with images and 15 minutes without images. Binary classifiers and confusion-prior baselines are trained using CPU in less than an hour in total. Results for runs with multiple seeds are shown in Table A1. Logistic regression, random forests and XGBoost [13] are trained using scikit-learn package [59] default parameters. For all models receiving user context that are not the transformer models, the lookback window is set to the full history. For transformer models, is is set to 50 questions. The knowledge tracing baselines are trained using the pyKT package [48] with default parameters. All baselines presented in Table A2 only predict binary outcomes, and receive user context.

Engineered features. For our binary classifiers and the simple MLP, context is encoded through additional features. For user context, these are a user's average accuracy, user's average accuracy on the image species, log-transformed counts of how often the user has seen the species in the past, and how many questions the user has answered on the same location. Additionally, boolean indicators are provided for whether the user is in their geographic focus region by Hex 3 location, their spatio-temporal focus by Hex 3 location and whether the user has improved in the past in any location over a 20 question sliding window. Species context is provided as average accuracies for the species and the choices. For our language models, user context is provided in form of tokenized history sequences. The history lookback window is set to 50 for our language models, and to the full history for all other models.

D Media Use

We used the following recordings from Cornell Lab of Ornithology | Macaulay Library: Fig. 2 uses ML614845753, ML624914011 and ML624836085. Fig. 4 upper row uses ML615927847, ML621578731, ML617550217 and ML621294128, lower row uses ML39633601, ML50619491, ML38293181 and ML226495281. Fig. 5 upper row uses ML30091521, ML117787821, ML302310521, ML83984151, ML141517111 and ML284199291, lower row uses ML51777001, ML26854421, ML301728521, ML290513131, ML50787721 and ML174404171. Fig. A3 uses ML463868861 and ML613090562.