EVALUATING THE ABILITY OF LARGE LANGUAGE MODELS TO GENERATE STRUCTURED CURRICULA FOR HUMAN LEARNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used to teach human learners, but it remains unclear what aspects of their instruction support learning. Most work has focused on interactive tutoring, where models can respond to questions, provide feedback, and adapt to a learner. Here, we isolate a complementary component: the curriculum a model chooses to present in advance. We evaluate LLM-generated curricula with 1,428 human learners across four concept-learning settings, progressing from controlled tasks in which teaching strategies can be characterized formally to more complex scientific and mathematical subjects. Where tractable, we compare LLM curricula with Bayesian pedagogy, a normative account in which teaching depends on selecting information with a learner's inference in mind. LLM-generated curricula align with the Bayes-optimal model, while exhibiting systematic preferences among distinct strategies that the theory treats as equally optimal, and these preferences benefit human learners. As the concepts become more compositional, models spontaneously introduce intermediate concepts and organize them into structured curricula that improve learning. The same organization appears in naturalistic subjects, where static LLM-generated curricula substantially improve human performance. Together, these results suggest that an important component of LLM teaching lies in how models select and structure instructional content, and that this organization extends from choosing examples to decomposing complex concepts into learnable parts for human learners.

1 INTRODUCTION

Large language models (LLMs) are increasingly being used as instructional systems for human learners. Recent work suggests that LLM-based tutors can improve learning outcomes (Kestin et al., 2025; Vanzo et al., 2025; Kumar et al., 2025) and provide academic support in settings with limited educational access (De Simone et al., 2025; Wang et al., 2025). Much of this work studies dialogue-based tutoring, where the model responds to a student's questions, provides explanations, and adapts through interaction. These affordances are important, but they leave open a more basic question about the instructional content itself: independent of dialogue, do LLMs choose useful concepts and examples, and do they organize them in ways that support human learning?

We study this question by evaluating curricula that models generate in advance. This removes dialogue, feedback, and online adaptation, and isolates decisions about what information to present and in what order. We use *curriculum* in the educational sense: a sequence of concepts and examples presented to a learner. Unlike curriculum learning in machine learning, where a fixed set of training examples is ordered for optimization, a curriculum here may also introduce intermediate concepts that were not part of the original target. Our goal is not to evaluate the full capabilities of an interactive tutor, but to identify whether the static structure of LLM-generated instruction contains a pedagogically useful strategy of its own and whether explicit prompting improves it further.

To characterize good instructional choices, we draw on Bayesian pedagogy (Shafto & Goodman, 2008; Shafto et al., 2014), which formalizes teaching as selective communication: the teacher anticipates what a learner will infer from the data they choose, while the learner in turn reasons about

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

why those particular data were selected for instruction. The teacher's selection therefore carries information beyond the observations themselves. In domains small enough to specify the learner's alternatives and the candidate data, the theory provides a normative account of pedagogical choice and predicts both the examples human teachers select and the inferences human learners make (Shafto et al., 2014). This gives us a setting in which the structure of good teaching can be characterized independently of whether a particular curriculum happens to produce a favorable outcome.

We first evaluate LLMs in two such domains: causal structure learning, where Bayesian pedagogy was originally validated against human teachers, and primitive visual concept learning. Across five models, LLM-generated examples fall within the Bayes-optimal set in 99.5–100% of primitive-task trials and closely track the Bayesian teacher's intervention choices in the causal task. Within the optimal set, however, LLMs favor contrastive pairs that isolate the defining feature directly; human learners are more accurate and substantially faster with these pairs than with the positive pairs favored by the Bayesian teacher. We refer to these two properties as *informative* and *legible*: the examples separate the target from its alternatives, while presenting the distinguishing information in a form that human learners can readily recover.

The same criterion becomes more difficult to apply when a concept is compositional. As the number of relevant parts grows, the alternatives multiply across those parts, explicit Bayesian computation becomes impractical, and a single example of the target may fail to isolate which component is diagnostic. In this regime, decomposition provides a natural extension of selective teaching: teach intermediate concepts that isolate reusable parts of the target, then recombine them. We test this first in a controlled compositional concept-learning task with approximately 17 million hypotheses, and then in three naturalistic subjects—logic circuits, z-scores, and Mendelian genetics—where the appropriate decomposition depends on domain structure and learners may bring prior knowledge.

Across these settings, we ask whether LLM-generated curricula exhibit the structure of pedagogical choice captured by Bayesian pedagogy, and whether that structure extends beyond the settings in which the theory can be computed.

2 Related Work

2.1 Language models as teachers of humans

Evidence that LLM-based instruction can improve human learning is accumulating. Randomized comparisons report gains from LLM tutors relative to conventional instructional conditions (Kestin et al., 2025), and systems that support human tutors can improve outcomes at scale (Wang et al., 2025). Other work has examined access to LLM-based educational support in settings where institutional resources are limited (De Simone et al., 2025). At the same time, the benefits of LLM assistance depend on how the system is used, and recent work suggests that instruction that substitutes for learner effort need not improve durable understanding (Bastani et al., 2025).

Most existing work treats the tutor as an interactive dialogue. Explanations for its effectiveness emphasize individualized pacing, immediate feedback, question answering, and adaptation to a learner's responses (Bloom, 1984; VanLehn, 2011; Kestin et al., 2025). Work on improving LLM tutors has similarly focused on the quality of tutor turns, dialogue policies, or pedagogical objectives defined over interactions (Dinucu-Jianu et al., 2025; Scarlatos et al., 2025), and recent evaluation frameworks score tutoring behavior against pedagogical rubrics (Macina et al., 2025; Scarlatos et al., 2025). These approaches address important components of tutoring, but they do not isolate the contribution of instructional selection and ordering independent of dialogue. We evaluate that component directly using static curricula and human learners.

2.2 Bayesian pedagogy and concept learning

Bayesian inference provides a unifying mathematical framework for learning as statistical updates over structured hypotheses (Tenenbaum et al., 2011). Bayesian pedagogy extends this framework to teaching by modeling teachers and learners as reasoning about one another's choices (Shafto & Goodman, 2008; Shafto et al., 2014). Formally, a rational learner updates their belief in hypothesis

h having observed some data d generated by the teacher according to Bayes' rule:

$$P_{\text{learner}}(h \mid d) = \frac{P_{\text{teacher}}(d \mid h)P(h)}{\sum_{h'} P_{\text{teacher}}(d \mid h')P(h')} \tag{1}$$

Simultaneously, the teacher considers what data to generate to maximize the learner's belief in the true hypothesis. However, direct maximization of $P_{\text{learner}}(h|d)$ is not robust when the teacher has an imperfect model of the learner. A soft maximization of the learner's posterior is used in practice:

$$P_{\text{teacher}}(d \mid h) \propto P_{\text{learner}}(h \mid d)^{\alpha} \tag{2}$$

where $\alpha \to \infty$ recovers direct maximization. Following Shafto et al. (2014), we use $\alpha = 1$ where data are chosen in direct proportion to the posterior probability they give to the true hypothesis. Equations 1 and 2 form a system of equations which are solved by fixed-point iteration.

In practice, Bayesian pedagogy predicts the choices made by both human teachers and learners (Shafto et al., 2014), and also captures other phenomena beyond conventional education settings, such as infant-directed speech (Eaves et al., 2016) and cooperative communication (Shafto et al., 2021). However, the framework has important limits. First, the hypothesis space, priors, and likelihoods require explicit specification (Griffiths et al., 2010), and each iteration of Equations 1–2 requires summing over every hypothesis and candidate example pair, which quickly becomes intractable as these spaces grow (Kwisthout et al., 2011; Goldman & Kearns, 1995). Prior work has compared LLMs to Bayesian teaching models in tractable domains (Ross & Andreas, 2024)—we extend this line of work to domains that grow increasingly complex and eventually exceed the scope of Bayesian pedagogy.

2.3 Decomposition and structured representations in language models

Recent post-training pipelines exploit the ability of LLMs to recognize the underlying skills and concepts for a given task to generate synthetic instruction-tuning data. These methods *decompose* a target concept into subconcepts, *organize* them into structured representations (e.g., knowledge trees), and *bootstrap* synthetic data that recombines those skills in novel ways.

Didolkar et al. (2024) demonstrate the usefulness of this approach for strengthening mathematical reasoning in LLMs, Li et al. (2024) to achieve broad performance gains across several domains, and Kaur et al. (2024) enumerate and recombine instruction-following skills to construct fine-tuning data. Across these approaches, the model-generated decomposition is useful because it exposes reusable structure that can be recombined during training.

Our use of decomposition is related in form but differs in purpose. We ask whether the intermediate concepts an LLM identifies are useful for teaching humans, whether they follow the structure of the target domain, and whether presenting them before the target compositional concept is what improves learning. We therefore treat decomposition as an observable property of the generated curriculum, and seek to test whether such decomposition is pedagogically useful.

3 Evaluating LLM pedagogy in tractable domains

We first examine two domains where Bayesian pedagogy is tractable, causal concept learning (Shafto et al., 2014) and primitive concept learning (Tenenbaum et al., 2011), asking whether LLM-generated instruction exhibits the structure of pedagogical choice captured by the Bayesian teacher and how LLMs resolve choices the model treats as equally optimal.

3.1 Causal concept learning

Behavioral task. Following Shafto et al. (2014), learners infer the structure of a latent Bayes net (Pearl, 2000) relating three binary variables. The 12 possible structures fall into three categories: common effect, causal chain, or common cause (Figure 1A). The teacher selects two interventions to inform the learner about a specified causal structure. Each intervention sets a variable *on*, after which its children activate with transmission rate 0.90; any variable also activates spontaneously with background rate 0.05.

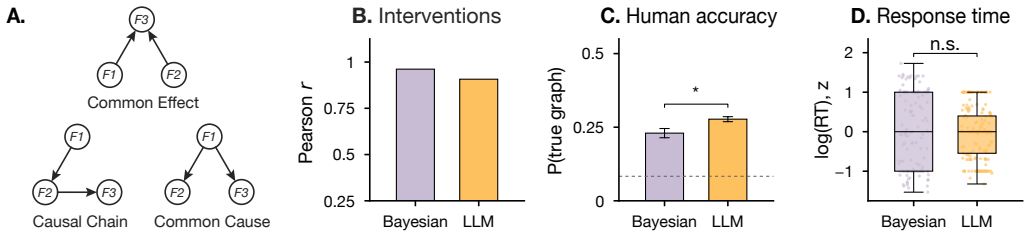


Figure 1: **Causal concept task results.** (A) Causal structures. (B) Correlation of Bayesian and LLM intervention distributions with the human distribution. (C) Human accuracy; dashed line marks chance ($1/12$). (D) Participant-normalized log response times. Error bars: SEM; boxes: IQR and median.

Curriculum generation. We queried each LLM in Table 1 for $n = 50$ intervention pairs per structure (600 per model). The Bayesian teacher's distribution over intervention pairs is taken from Shafto et al. (2014), and weights each pair according to how strongly it distinguishes the target structure from alternatives for a learner who assumes that the pair was chosen for teaching. Prompts are in Appendix B.4.

Human evaluation. $N = 217$ participants from Prolific (Prolific, 2026) completed ten trials each. On each trial, the intervention pair was drawn from one of six sources (five LLMs or the Bayesian model), and participants rated each of the 12 candidate structures from 0 to 100 (Shafto et al., 2014). Details are in Appendix B.

3.2 Causal experiment results

LLM interventions align with Bayesian pedagogy. LLM intervention distributions correlated strongly with the Bayesian distribution on common effect ($r \in [0.88, 0.91]$) and causal chain ($r \in [0.85, 0.95]$) structures across all five models, and even more strongly with the human distribution (Shafto et al., 2014). Common cause structures showed greater variability ($r \in [-0.00, 0.82]$; Table 2, Appendix B.3).

Concentration of intervention distributions. LLM and Bayesian distributions differ in their concentration. The Bayesian and human distributions place non-trivial probability mass across multiple intervention pairs (mean entropy $H = 2.22$ and 1.69 bits respectively), while LLMs concentrate sharply on one or two pairs (mean $H \in [0.26, 0.65]$ across models). Full LLM intervention distributions, per-cell entropy values, and human-distribution comparisons are reported in Appendix B.

Behavioral results. LLM curricula yielded higher learner accuracy than Bayesian curricula for common effect (mean normalized probability on the true graph 0.31 vs. 0.24; $\hat{\beta} = +0.090$, $p = 0.026$) and comparable accuracy for causal chain and common cause ($p \geq 0.60$). The pooled mixed-effects contrast was $\hat{\beta} = +0.034$ ($p = 0.047$); a logistic GLMM on a stricter accuracy criterion (true graph rated as the unique maximum) yielded a directionally consistent but non-significant LLM advantage ($p = 0.12$). Although LLMs concentrate on a single intervention pair while the Bayesian model spreads probability across several, learner accuracy is largely comparable. LLMs tend to select the modal pair under the Bayesian distribution, indicating that their more concentrated choices remain highly informative for human learners. Full results are in Appendix B.

3.3 Primitive concept learning

Behavioral task. We next test whether the ability to select informative examples extends to learning a rule based on a single visual feature. The teacher selects two stimuli and labels each as a positive or negative instance of the target concept (Figure 2A). Each concept maps a pseudolanguage label to a single visual feature of a specific body region (e.g., *dax* is objects with red bodies). The learner observes the labeled pair and selects the correct rule from the full hypothesis space (Appendix C.2).

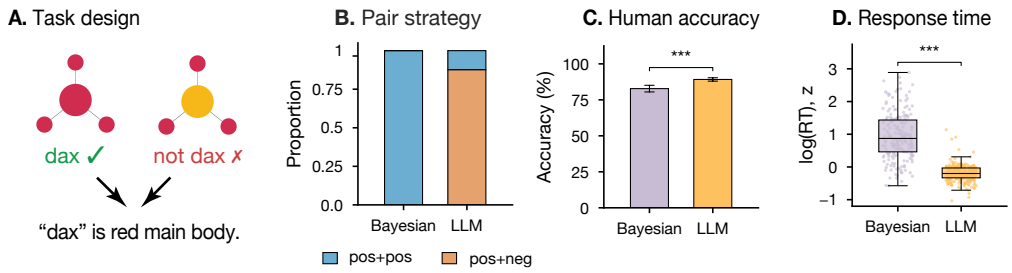


Figure 2: **Primitive concept task results.** (A) Example stimuli and concept. (B) Proportion of contrastive and positive pairs. (C) Human accuracy. (D) Participant-normalized log response times (correct trials). Error bars: SEM; boxes: IQR and median.

Curriculum generation. We followed the same curriculum-generation procedure as in the causal task. The Bayesian pedagogical model identified two classes of maximally informative pairs: *contrastive*, in which stimuli differ only on the target feature, and *positive* pairs, in which stimuli share only the target feature. Although both are equivalently informative, the positive pairs admit more degrees of freedom while remaining optimal, so the model places nearly all probability mass on them ($\approx 99.8\%$ across concepts).

Human evaluation. We recruited $N = 283$ participants from Prolific. Participants were presented with two labeled stimuli per trial for ten trials and selected their answer for each trial via a drop-down interface. Details are in Appendix C.

3.4 Primitive experiment results

LLMs disambiguate optimally but differ in strategy from Bayesian inference. Across 2,994 trials, LLMs produced Bayes-optimal example pairs in 99.5%–100% of cases. LLMs and the Bayesian model nonetheless prefer distinct strategies among the optimal options: the Bayesian model concentrates on positive pairs, while LLMs prefer contrastive pairs (87.5%–99.7%; Appendix C).

Optimality of diverging teaching strategies. The two strategies are Bayes-optimal only when executed in different ways. A contrastive pair must hold every irrelevant dimension fixed so that the target feature is the only difference; a positive pair must vary the irrelevant dimensions so that the target feature is the only stable commonality. LLMs execute both strategies in these optimal forms. Contrastive pairs differ on *zero* of the 11 unconstrained feature dimensions in 99.5%–100% of trials, while positive pairs differ on *all* 11 unconstrained dimensions in 100% of trials. Uniform random sampling would produce neither pattern (Appendix C). Thus, LLMs do not simply produce pairs that happen to fall within the optimal set; their examples isolate the target feature in the manner required by the chosen teaching strategy.

Behavioral experiment results. Participants (2,830 trials) identified target concepts more accurately from LLM curricula than from Bayesian curricula (85.4% vs. 80.6%, $p < 0.05$; Figure 2C), and were also substantially faster (median response time 13s vs. 23s; $p < 0.001$; Figure 2D; mixed-effects models with random participant and concept intercepts). Together, these results indicate that the contrastive strategy preferred by LLMs is more pedagogically effective for learners: contrastive pairs allow learners to isolate the defining feature through direct comparison, whereas the positive pairs preferred by the Bayesian model require learners to hold many feature differences in working memory simultaneously. The Bayesian objective treats the two strategies as equally informative, but it does not represent the working-memory costs that differ between them. Human learners are more accurate and substantially faster with the contrastive pairs preferred by LLMs. Within the set of examples that are normatively informative, LLMs therefore favor a strategy that is more legible to human learners.

3.5 Summary of findings with causal and primitive concepts

Across both paradigms, LLMs and the Bayesian teacher agree on which examples are informative, while LLMs exhibit an additional preference over how that information is presented. In the causal

task, LLMs concentrate on a small subset of highly diagnostic interventions and match the learner accuracy of the full Bayesian distribution; in the primitive task, they systematically prefer minimal contrasts, which yield higher accuracy and faster responses. LLM-generated curricula are thus both informative under the normative criterion and, within that set, sensitive to how readily human learners can recover the relevant distinction.

4 COMPOSITIONAL CONCEPT LEARNING

The primitive and causal paradigms were small enough that the alternatives can be enumerated and the distinguishing information can be isolated in a small number of examples, enabling explicit Bayesian modeling. However, efficient learning does not require the overall concept space to be small; it requires that the relevant distinction be isolable. For a compositional concept that is intractable for Bayesian pedagogy and cannot be merely taught with a limited set of examples, this can be achieved by introducing intermediate concepts that isolate reusable parts before recombining them into the target. We next test whether LLMs generate such decompositions when provided the option, and whether the resulting curricula improve human learning.

4.1 METHODS

Behavioral task. Participants learned pseudolanguage labels for visual concepts defined over conjunctions of object features (Figure 3A). The learner's goal is to infer the underlying rule for each target concept after receiving an LLM-generated curriculum that either introduces intermediate subconcepts (decomposed condition) or presents examples of the target concepts alone (target-only condition).

Bayesian intractability. For a compositional concept defined over k feature dimensions, both the hypothesis space and the number of examples needed to distinguish the target from similar but distinct alternatives grows exponentially. This scaling makes Bayesian pedagogical inference impractical (≈ 17 million hypotheses; calculations in Appendix D.2).

Curriculum generation. For each set of target concepts, we prompted each LLM in Table 1 with a two-stage protocol ($n = 50$ curricula per model). In the first stage, we provided LLMs with a set of target concepts and asked for a curriculum that would help a student learn the targets. For each curriculum step, the model specified the concept name, definition, and relevant prerequisite concepts. In the second stage, we prompted the same model to generate four labeled examples per curriculum step. In the target-only condition, each LLM generated four illustrative examples for each target concept without introducing subconcepts ($n = 50$ per model).

Human evaluation. We recruited $N = 383$ participants from Prolific, randomly assigned to the decomposed or target-only condition. After completing the curriculum, participants identified each target concept's underlying rule using a drop-down interface. Details are in Appendix D.

4.2 RESULTS

LLMs effectively decompose target concepts. Although the prompt did not require the model to introduce subconcepts, specify their number, or prescribe a decomposition strategy, LLMs consistently introduced subconcepts before teaching the target concepts. Across 906 non-target curriculum steps, no model introduced single-feature primitive concepts. Instead, 54%–100% of non-target steps were chunked according to different spatial locations of the concepts (Appendix D). These results suggest that LLMs preferred to decompose compositional concepts into reusable instructional "chunks" that preserved the spatial structure of the stimulus domain. That no model introduced single-feature sub-concepts to a single body part indicates that the models organized instruction around reusable parts of the stimulus structure rather than around arbitrary feature fragments.

Decomposed curricula improved compositional learning. The most common strategy was *flat decomposition* ($\approx 61\%$ of curricula; Figure 3B), in which subconcepts were taught as standalone building blocks and later recombined at the target steps. The second most common was *progressive decomposition* ($\approx 29\%$), which introduced back-to-back dependencies between subconcepts. GPT-5.2 was distinctively willing to nest subconcepts *hierarchically* (34% of its curricula, vs. $\leq 6\%$ for other models). Per-model breakdowns are in Appendix D.5.

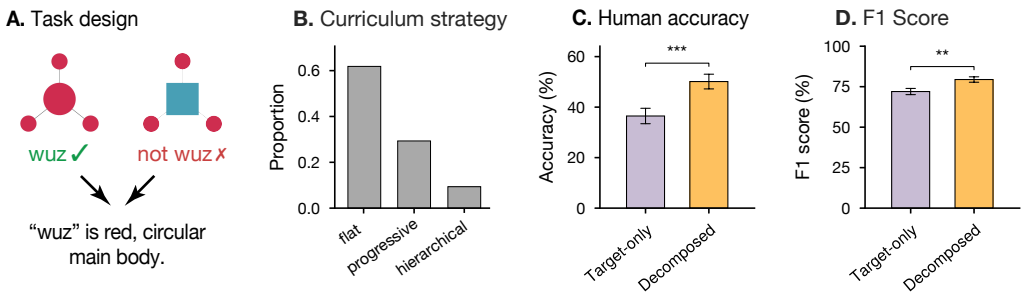


Figure 3: **Compositional concept task results.** (A) Task schematic. (B) Proportion of flat, progressive, and hierarchical curricula. (C) Human accuracy and (D) feature-level F1 by condition. Error bars: SEM.

These decomposed curricula improved human learning relative to the target-only baseline (Figure 3C,D). In our preregistered mixed-effects logistic regression, participants were more accurate at reconstructing the composite rule after learning from decomposed curricula than from the target examples alone (50.1% vs. 36.5%; $\hat{\beta} = 1.41$, $p < 0.001$). Subconcept performance also correlated with target performance (Spearman $\rho = 0.60$, $p < 10^{-20}$), consistent with intermediate concepts contributing to target understanding. This benefit was not explained by additional practice with the interface (Appendix D.8).

Of the two most common strategies, flat decomposition significantly improved performance relative to the target-only baseline ($\Delta = +16.3$ pp; $p < 0.001$); progressive decomposition trended in the same direction but did not reach significance ($\Delta = +7.5$ pp; $p = 0.19$; Table 8).

5 Naturalistic concept learning

The compositional experiment shows that LLMs can improve learning by introducing intermediate concepts before a complex target. Real instructional subjects have the same broad structure: predicting the output of a multi-gate circuit, computing a probability from a z-score, or predicting offspring genotypes each requires constituent ideas that are typically introduced first. Unlike preceding experiments, learners may already know some of those constituents. We therefore use three naturalistic subjects to test whether LLM-generated curricula remain effective when the target structure is less constrained and prior knowledge may be present.

This setting also allows us to distinguish explanations for the benefit of decomposition. If decomposition is part of the model's default teaching strategy, single-pass curricula should already contain prerequisite structure and removing the preceding modules should reduce performance; an explicit decomposition stage should add little beyond the structure the model supplies by default. A no-curriculum condition measures baseline knowledge in each subject.

5.1 Methods

Subjects. We selected three topics (logic circuits, z-scores, and Mendelian genetics) from introductory scientific and mathematical instruction (College Board, 2020b;c;a) whose target skills combine multiple constituent concepts. For instance, in Mendelian genetics, predicting offspring genotype ratios requires concepts including alleles, dominance, gamete formation, and Punnett-square reasoning.

Behavioral task. We conduct two experiments in which participants were assessed on their knowledge of one of the three domains listed above in an LLM-generated multiple choice exam. The exam consists of twelve questions and is identical across all models and conditions within a domain.

Curriculum generation. Four conditions form an ablation gradient. In the *two-stage* condition, the model first decomposes the learning objective and then writes a curriculum from that decomposition. In the *single-pass* condition, it writes the curriculum in one prompt without an explicit decomposition step. In the *target-only* condition, participants receive only the final module of a curriculum,

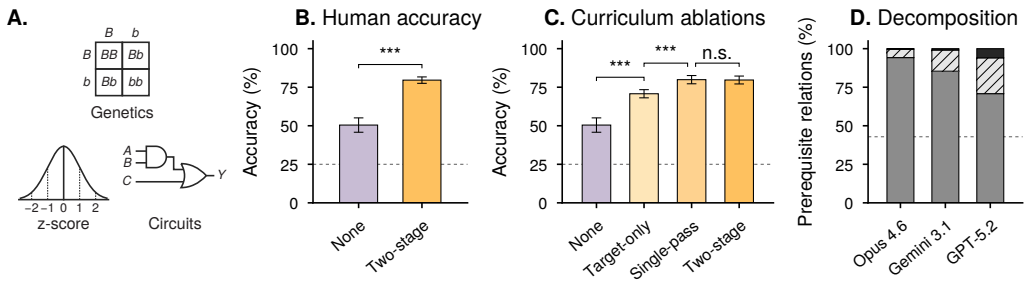


Figure 4: **Naturalistic concept task results.** (A) Domains. (B) Accuracy with and without two-stage curricula (Experiment 1). (C) Accuracy across curriculum ablations (Experiment 2). (D) Prerequisite order in single-pass curricula relative to each model's decomposition plan: taught earlier (gray), in the same unit (hatched), or later (black). Dashed lines: chance (B–C) and random ordering (D). Error bars: SEM.

and in the *no curriculum* condition, participants complete the assessment without instruction. Adjacent contrasts isolate explicit decomposition (two-stage vs. single-pass), the preceding modules (single-pass vs. target-only), and the curriculum itself (target-only vs. no curriculum).

Experiments. In the first experiment, participants either received no curriculum or studied a two-stage curriculum generated by Claude Opus 4.6, Gemini 3.1 Pro Preview, or GPT-5.2. In the second, participants studied a target-only, single-pass, or two-stage curriculum. Because the first experiment showed no detectable differences in learning across curriculum-generating models, all curricula in the second experiment were generated by Gemini 3.1 Pro Preview. The no-curriculum group from the first experiment serves as the baseline for both analyses. For each model, condition, and subject, we generated three curriculum sets and randomly assigned participants to a set.

Human evaluation. We recruited $N = 222$ participants for the first experiment and $N = 323$ for the second from Prolific, and randomly assigned each participant to one curriculum condition and subject. Details are in Appendix E.

5.2 Results

LLM-generated curricula improve learning in naturalistic subjects. Participants who studied a two-stage curriculum answered 79.6% of exam items correctly, compared with 50.4% without a curriculum (chance = 25%; $\hat{\beta} = 1.31$, $p < 0.001$; Figure 4B). The three models produced similar learner performance (Claude Opus 4.6: 79.5%; Gemini 3.1 Pro Preview: 79.4%; GPT-5.2: 79.9%), and the variance of the preregistered random intercept for LLM was estimated at zero (Table 11). No-curriculum performance differed across subjects (logic circuits: 28.1%; z-scores: 56.0%; Mendelian genetics: 67.2%), indicating differences in prior knowledge, and the curriculum improved performance in each (to 82.6%, 78.9%, and 77.4%). These results extend the compositional finding to instructional domains with familiar content and nonzero learner priors.

Models implicitly decompose their curricula, facilitating learner understanding. In the second experiment, the target-only curriculum outperformed no curriculum (70.8% vs. 50.4%; $\hat{\beta} = +0.79$, $p < 0.001$), and the single-pass curriculum outperformed target-only (79.9% vs. 70.8%; $\hat{\beta} = +0.53$, $p < 0.001$), but two-stage and single-pass curricula were comparably effective (79.7% vs. 79.9%; $\hat{\beta} = -0.06$, $p = 0.61$). This null result is consistent with the structure of single-pass curricula—using each model's two-stage decomposition plans as a reference for prerequisite relations, freely-generated single-pass curricula from all three models taught the prerequisite in an earlier instructional unit than its dependent concept for 83.3% of relations, compared with 42.9% expected under random ordering ($p < 0.001$; Figure 4D). These results indicate that the implicit decompositions of the single-pass curricula are consistent with those generated by explicit instruction in the two-stage condition.

6 DISCUSSION

Across four concept-learning settings, we evaluated whether LLM-generated curricula exhibit a coherent pedagogical strategy: first where good teaching choices can be characterized formally, and then where teaching requires structuring increasingly complex material. The results show a consistent progression. In tractable domains, LLM choices closely track the teaching strategies identified by Bayesian pedagogy, while also showing systematic preferences among choices the theory treats as equivalent. As the concepts become compositional, the models introduce intermediate concepts and organize them before the target, and this structure supports learning in both controlled and naturalistic domains.

The first experiments therefore show that LLMs reproduce much of the structure that normative Bayesian methods predict. At the same time, the primitive task exposes a limit of the normative model for predicting human learning: when two strategies are equally informative under the theory, LLMs strongly prefer the contrastive strategy, and human learners recover the concept more accurately and faster from it. As such, the models' choices are not only informative in the sense captured by the theory, they also favor forms of presentation that are easier for human learners to use. We describe this additional property as *legibility*, without claiming that the models explicitly represent human cognitive constraints or perform Bayesian inference.

This distinction becomes important as the teaching problem grows. For a compositional target, the teacher must decide not only which examples to present, but which intermediate concepts should be introduced before the target can be understood. LLMs make this broader pedagogical choice by decomposing targets into reusable concepts that follow the structure of the domain. The naturalistic experiments extend the same pattern to scientific and mathematical subjects, where the relevant prerequisite structure must be inferred rather than specified in advance. We find that LLM-generated curricula contain implicit decompositions of the target domain that are consistent with the decompositions produced by models upon explicit instruction. Accordingly, within the same model and for the domains that we test, curricula generated within a single pass are comparable to curricula that receive the knowledge hierarchy within the prompt. In either case, the curricula were modular, and removing preceding modules eliminated roughly a third of the learning gain over no curriculum. In this sense, decomposition is not a separate capability from the example selection studied earlier, but an extension of learner-directed pedagogical organization to a richer action space.

Finally, the static-curriculum design isolates a component of LLM teaching that does not depend on dialogue, feedback, or online adaptation. This complements work on interactive tutoring: what a model chooses to present, and how it organizes that material, can itself support learning. It also suggests a level at which LLM tutors can be evaluated beyond outcomes or individual dialogue turns. Where Bayesian pedagogy is tractable, example choices can be compared with a normative standard; in more complex domains, curriculum structure can be evaluated through the prerequisites a model introduces, their granularity, and their ordering. Interactive systems may add adaptation on top of this structure by changing what is introduced according to what a learner already knows.

Limitations Our experiments isolate selection and organization within static curricula, and therefore do not evaluate explanation, questioning, feedback, or adaptation. The naturalistic extension remains limited to three topics, short sessions, and adult online participants. The target-only ablation also removes instructional content as well as prerequisite structure, so dosage remains a possible contributor. Finally, the naturalistic ablation uses a single curriculum-generating model — replication across models, learner populations, and delayed assessments will be important for establishing generality.

Conclusion. While LLMs have been the focus of interactive, personalized education systems, much less is known about how they select and structure pedagogical content. We generate LLM curricula across increasingly complex learning paradigms and evaluate their efficacy for teaching human learners. We find that LLMs converge on a subset of Bayes-optimal choices in tractable settings, which teaches learners as well as or better than the Bayesian pedagogical distribution. We demonstrate that LLMs produce similar decompositions regardless of whether they are explicitly prompted to decompose, and that their decompositions support learning. Together, these results point towards example selection and decomposition as central abilities supporting LLM pedagogy.

AI USE STATEMENT

In this work, the authors used generative AI tools for editing code and to perform sentence-level editing of the manuscript. The authors checked AI-assisted code and its outputs, and responsibility for the final content of this work including text, claims or artifacts produced with the aid of generative AI. Furthermore, the authors used generative-AI tools for creating the curricula evaluated in this study.

ETHICS STATEMENT

This work involves research with human participants. The study was conducted under a protocol approved by an institutional review board (IRB). All participants provided informed consent before participation and were informed that participation was voluntary and that they could withdraw at any time. Responses and reaction times were recorded while participants completed their tasks, as described in the appendix.

REPRODUCIBILITY STATEMENT

All models used to generate curricula and their reasoning settings are listed in Table 1. Prompting protocols for each task are also provided in the appendix (Appendices B.4, C.5, D.9, and E.4). Participant recruitment, task interfaces, exclusion criteria, and the preregistered analyses for each experiment are described in Appendices B–E. Pre-registrations will be linked after review to preserve anonymity. Because LLM outputs vary across calls, we will release all generated curricula after the review period, along with de-identified participant data and analysis code.

REFERENCES

Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122, 2025. doi: 10.1073/pnas.2422633122.

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

College Board. AP biology course and exam description. `https://apcentral.collegeboard.org`, 2020a.

College Board. AP computer science principles course and exam description. `https://apcentral.collegeboard.org`, 2020b.

College Board. AP statistics course and exam description. `https://apcentral.collegeboard.org`, 2020c.

Martín De Simone, Federico Tiberti, Maria Barron Rodriguez, Federico Manolio, Wuraola Mosuro, and Eliot Jolomi Dikoru. From chalkboards to chatbots: Evaluating the impact of generative AI on learning outcomes in Nigeria. Policy Research Working Paper 11125, World Bank, 2025.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. Metacognitive capabilities of LLMs: An exploration in mathematical problem solving. In *Advances in Neural Information Processing Systems*, 2024.

David Dinucu-Jianu, Jakub Macina, Nico Daheim, Ido Hakimi, Iryna Gurevych, and Mrinmaya Sachan. From problem-solving to teaching problem-solving: Aligning llms with pedagogy using reinforcement learning, 2025. URL `https://arxiv.org/abs/2505.15607`.

Baxter S. Eaves, Naomi H. Feldman, Thomas L. Griffiths, and Patrick Shafto. Infant-directed speech is consistent with teaching. *Psychological Review*, 123(6):758–771, 2016. doi: 10.1037/rev0000031.

Sally A Goldman and Michael J Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995.

Thomas L Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B Tenenbaum. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, 14(8):357–364, 2010.

Peter Harrison, Raja Marjieh, Federico Adolfi, Pol van Rijn, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. Gibbs sampling with people. *Advances in Neural Information Processing Systems*, 33:10659–10671, 2020.

Simran Kaur, Simon Park, Anirudh Goyal, and Sanjeev Arora. Instruct-SkillMix: A powerful pipeline for LLM instruction tuning. *arXiv preprint arXiv:2408.14774*, 2024.

Greg Kestin, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti. AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 2025.

Harsh Kumar, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. Math education with large language models: Peril or promise? In *Artificial Intelligence in Education (AIED)*, volume 15880 of *Lecture Notes in Computer Science*. Springer, 2025. doi: 10.1007/978-3-031-98459-4_5.

Johan Kwisthout, Todd Wareham, and Iris Van Rooij. Bayesian intractability is not an ailment that approximation can cure. *Cogn. Sci.*, 35(5):779–784, 2011.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, and Furu Wei. Synthetic data (almost) from scratch: Generalized instruction tuning for language models, 2024. URL https://arxiv.org/abs/2402.13064.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. *arXiv preprint arXiv:2502.18940*, 2025.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2000.

Prolific. Prolific Online participant recruitment platform. https://www.prolific.com, 2026. Accessed: 2026.

Alexis Ross and Jacob Andreas. Toward in-context teaching: Adapting examples to students' misconceptions. *arXiv preprint arXiv:2405.04495*, 2024.

Alexander Scarlatos, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. *Training LLM-Based Tutors to Improve Student Learning Outcomes in Dialogues*, pp. 251–266. Springer Nature Switzerland, 2025. ISBN 9783031984143. doi: 10.1007/978-3-031-98414-3_18. URL http://dx.doi.org/10.1007/978-3-031-98414-3_18.

Patrick Shafto and Noah D. Goodman. Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2008.

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

Patrick Shafto, Junqi Wang, and Pei Wang. Cooperative communication as belief transport. *Trends in Cognitive Sciences*, 25(10):826–828, 2021. doi: 10.1016/j.tics.2021.07.012.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.

Kurt VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, 46(4):197–221, 2011.

Alessandro Vanzo, Sankalan Pal Chowdhury, and Mrinmaya Sachan. GPT-4 as a homework tutor can improve student engagement and learning outcomes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Rose E. Wang, Ana T. Ribeiro, Carly D. Robinson, Susanna Loeb, and Dora Demszky. Tutor copilot: A human-ai approach for scaling real-time expertise, 2025. URL https://arxiv.org/abs/2410.03017.

A General Experimental Details

A.1 Models and Querying Protocol

We evaluate five frontier large language models from three providers (Table 1). All models are queried through their respective providers' APIs with temperature 1.0 and `max_completion_tokens` 16,384.

Table 1: LLMs and reasoning configurations per task. For Anthropic and OpenAI models, settings refer to the qualitative reasoning effort level; for Google models, to the maximum number of thinking tokens.

Model	Provider	Causal	Primitive	Compositional
Claude Opus 4.6	Anthropic	medium	high	high
Claude Sonnet 4.6	Anthropic	medium	high	high
GPT-5.2	OpenAI	medium	high	high
Gemini 3.1 Pro Preview	Google	8,192	32,768	24,576
Gemini 3 Flash Preview	Google	8,192	32,768	24,576

Prompting protocol. Across all experiments, we first prompt models to formulate their curricula and explain their pedagogical reasoning. We then pass this free-form response to a cheaper formatting model (GPT-4o mini) for parsing into structured JSON. This two-stage scheme produces qualitatively better curricula than directly prompting for JSON output, and preserves an interpretable intermediate output. Outputs failing schema validation are discarded. Full prompts appear in the task-specific sections below.

Deployment per participant. In the primitive and compositional experiments, each participant receives a unique curriculum generated specifically for them, reflecting realistic LLM deployment conditions. The causal experiment uses a different approach (Appendix B.2): we approximate a pedagogical distribution via repeated LLM queries and resample intervention pairs from it, rather than assigning individual LLM outputs directly.

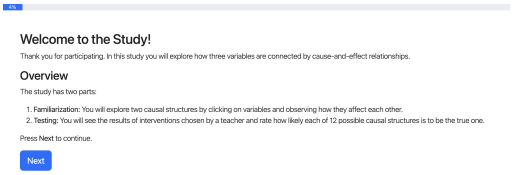
A.2 Participants

Participants were recruited from Prolific (US residents, native English speakers, normal or corrected-to-normal vision) and compensated at $15.32–$19.53 per hour. Per pre-registration, we excluded individual trials with reaction times below 1 second or more than 3 SD above the participant's mean, and excluded participants with chance-level overall accuracy. All studies were approved by the authors' institutional IRB [details withheld for anonymous review]; informed consent was obtained before each session. Full pre-registered protocols are available on AsPredicted and will be provided after review.

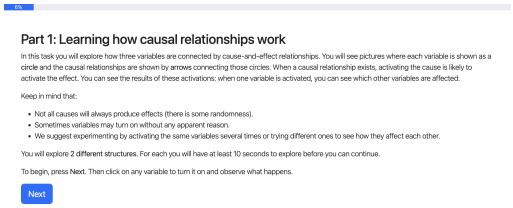
B CAUSAL CONCEPT LEARNING

B.1 TASK INTERFACE

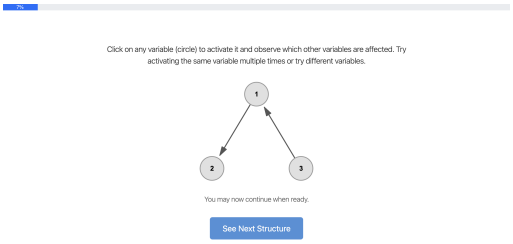
Data were collected using PsyNet (Harrison et al., 2020) to obtain 30 valid responses per unique curriculum condition. Figure 5 walks through the full participant experience, from landing page to the main rating screen. Participants are first introduced to the three binary variables and the notion of a probabilistic intervention, then proceed to the trial interface, where they see two pictures depicting the selected interventions and resulting variable states and rate all 12 candidate causal structures on a 0–100 scale.



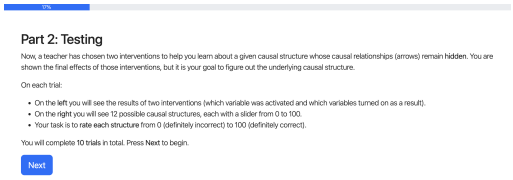
(a) Welcome screen with study overview and informed consent.



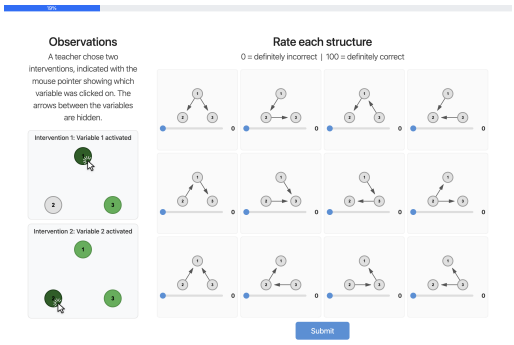
(b) Initial instructions introducing the three binary variables and the cover story.



(c) Explanation of interventions and the probabilistic nature of causal links.



(d) Rating instructions describing the 0–100 confidence scale used on each trial.



(e) Main trial interface showing the two intervention outcomes and the 12 candidate causal structures to be rated.

Figure 5: Participant interface for the causal concept learning experiment, presented in sequence: (a) welcome, (b) task setup, (c) intervention mechanics, (d) rating instructions, and (e) the trial interface.

B.2 Curriculum Generation

Following Shafto et al. (2014), we randomly permute variable labels on each causal structure query before prompting LLMs to control for orientation bias. For each model–structure pair, we aggregate responses across $n = 50$ queries to estimate an empirical distribution over intervention pairs, then sample from this distribution to assign curricula to participants. This resampling step further controls for orientation effects, as there is no natural mapping from named nodes in the prompt to visual positions on the participant interface. Bayesian pedagogical curricula are generated by sampling from the theoretical distributions of Shafto et al. (2014) using the same procedure. The Bayesian pedagogical distribution weights intervention pairs by their informativeness to a rational learner who reasons optimally about the teacher's intent.

B.3 Full Results

Table 2 reports, for each model and causal structure, the Pearson correlation between the LLM's empirical intervention distribution and both the Bayesian and human pedagogical distributions (Shafto et al., 2014), alongside Shannon entropy of the LLM distribution. Bootstrap 95% CIs use $B = 5{,}000$ resamples; Bayesian and human values are fixed references. Entropy is bounded in $[0, \log_2 6] = [0, 2.585]$, with $H = 0$ indicating a point-mass distribution and $H = 2.585$ the uniform distribution.

Among common cause structures, Claude Opus 4.6 and GPT-5.2 maintain strong correlation with the Bayesian distribution ($r = 0.78$ and 0.82), Claude Sonnet 4.6 and Gemini 3.1 Pro Preview show weaker correlation ($r = 0.51$ and 0.47), and Gemini 3 Flash Preview shows no detectable correlation ($r = -0.004$, 95% CI $[-0.10, 0.10]$).

Table 2: Pearson correlations of LLM intervention distributions against the Bayesian and human pedagogical distributions (Shafto et al., 2014), and Shannon entropy of LLM distributions. Entropy CIs are bootstrap ($B = 5{,}000$). Reference entropies are averaged across structures for brevity (per-structure Bayesian values: 2.255 / 2.212 / 2.198; human: 1.727 / 1.616 / 1.714).

		Pearson r		LLM entropy (bits)	
Model	Structure	vs. Bayesian	vs. Human	H	95% CI
Reference (mean across structures)					
	Bayesian	—	—	2.22	—
	Human	—	—	1.69	—
Claude Opus 4.6	Common effect	0.91	0.98	0.05	$[0.00, 0.11]$
	Causal chain	0.95	0.97	0.13	$[0.00, 0.25]$
	Common cause	0.78	0.90	1.31	$[1.17, 1.42]$
Claude Sonnet 4.6	Common effect	0.88	0.97	0.58	$[0.38, 0.76]$
	Causal chain	0.85	0.86	1.11	$[0.98, 1.22]$
	Common cause	0.51	0.67	0.05	$[0.00, 0.11]$
GPT-5.2	Common effect	0.91	0.98	0.05	$[0.00, 0.11]$
	Causal chain	0.95	0.97	0.25	$[0.09, 0.39]$
	Common cause	0.82	0.94	1.65	$[1.51, 1.75]$
Gemini 3.1 Pro Preview	Common effect	0.91	0.98	0.00	$[0.00, 0.00]$
	Causal chain	0.95	0.97	0.19	$[0.08, 0.31]$
	Common cause	0.47	0.64	0.58	$[0.41, 0.72]$
Gemini 3 Flash Preview	Common effect	0.91	0.98	0.00	$[0.00, 0.00]$
	Causal chain	0.95	0.97	0.00	$[0.00, 0.00]$
	Common cause	0.00	0.15	1.17	$[1.06, 1.26]$

B.3.1 Per-Pair Intervention Distributions

B.3.2 Per-Model Behavioral Results

Table 3 disaggregates the causal behavioral contrast (Figure 1C) into per-model means and per-(model, structure) contrasts vs. the Bayesian baseline. Each contrast is estimated from a linear mixed-effects model on `p_correct` (rating on the true graph divided by the sum of all 12 ratings)

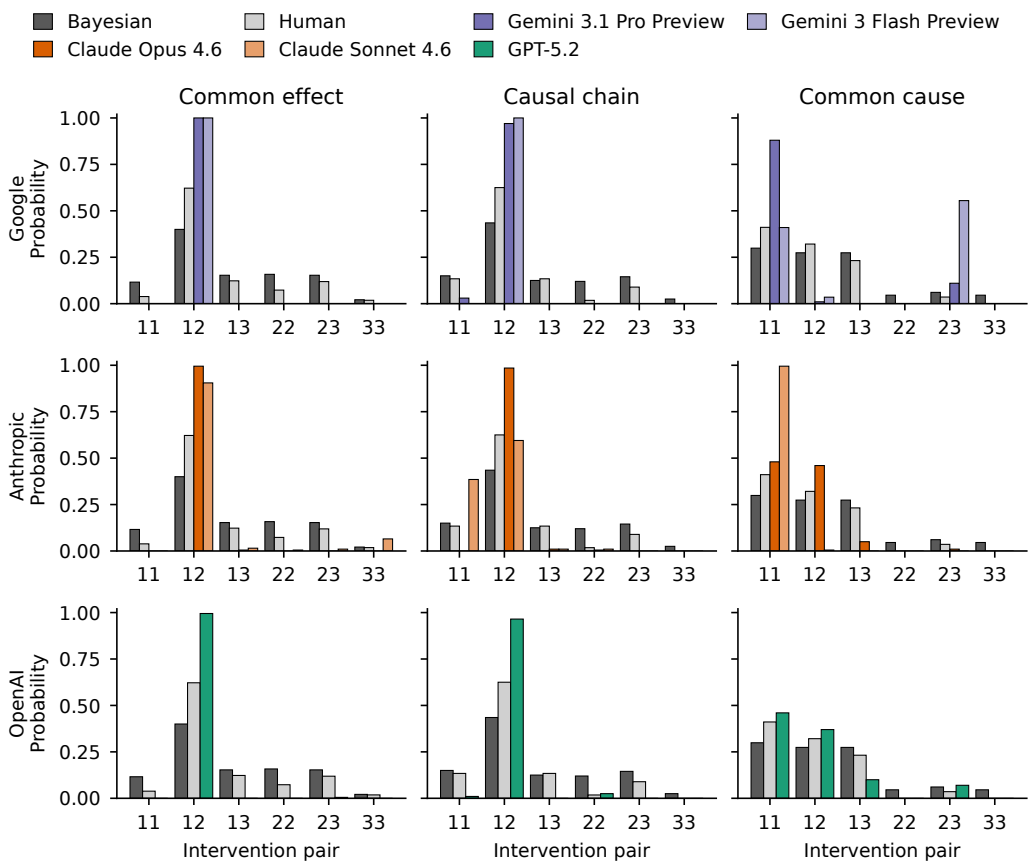


Figure 6: Full empirical distribution over the six intervention pairs for each (model, structure) cell, alongside the Bayesian and human distributions from Shafto et al. (2014).

with random participant intercepts, restricted to trials sourced from that model or the Bayesian baseline within the relevant structure.

Table 3: Per-(model, structure) rating accuracy on the true graph and contrast vs. the Bayesian baseline. *Means* report mean `p_correct`. $\hat{\beta}$ is the LLM vs. Bayesian fixed effect from a linear mixed-effects model with a random participant intercept, fit separately within each (model, structure) cell.

Model	Structure	LLM mean	Bayes mean	$\hat{\beta}$	SE	t	p
Claude Opus 4.6	common effect	0.324	0.240	$+0.066$	0.052	$+1.27$	0.206
	causal chain	0.220	0.188	$+0.004$	0.029	$+0.15$	0.882
	common cause	0.289	0.300	-0.012	0.046	-0.26	0.793
Claude Sonnet 4.6	common effect	0.325	0.240	$+0.093$	0.055	$+1.69$	0.094
	causal chain	0.235	0.188	$+0.028$	0.029	$+0.98$	0.326
	common cause	0.344	0.300	$+0.042$	0.048	$+0.88$	0.380
GPT-5.2	common effect	0.340	0.240	$+0.113$	0.051	$+2.23$	0.027
	causal chain	0.261	0.188	$+0.055$	0.026	$+2.10$	0.037
	common cause	0.297	0.300	-0.017	0.046	-0.36	0.716
Gemini 3 Flash Preview	common effect	0.315	0.240	$+0.100$	0.047	$+2.14$	0.034
	causal chain	0.267	0.188	$+0.043$	0.029	$+1.47$	0.143
	common cause	0.273	0.300	-0.027	0.047	-0.57	0.572
Gemini 3.1 Pro Preview	common effect	0.270	0.240	$+0.026$	0.045	$+0.57$	0.570
	causal chain	0.243	0.188	$+0.030$	0.030	$+1.00$	0.319
	common cause	0.312	0.300	$+0.016$	0.044	$+0.36$	0.719

Pooled model and pre-registered LME. Following our pre-registration, the pooled model fit to $n = 1{,}982$ trials from 205 participants (excluding 40 trials on which a participant gave all 12 candidate graphs a rating of 0, leaving `p_correct` undefined) is

$$\texttt{p_correct} \sim \texttt{source_c} + (1 \mid \texttt{participant}) + (1 \mid \texttt{structure}),$$

where `source_c` is the LLM vs. Bayesian contrast (LLM = $+0.5$, Bayesian = -0.5). The fixed effect is $\hat{\beta} = +0.034$, SE $= 0.017$, $t(1827.5) = 1.99$, $p = 0.047$. A logistic GLMM on the binary outcome "true graph rated as the unique maximum" (same random structure) on the full sample size of $n = 2{,}022$ yields pooled OR $= 1.28$, 95% CI $[0.94, 1.73]$, $p = 0.12$.

B.4 Prompt

```
You are an expert in concept learning.
Your task is to help a learner infer the correct underlying latent causal
     structure of a system involving three binary variables: f1, f2, and
    f3.
You have full knowledge of the true structure:

Structure: f2 -> f3 -> f1
f1 does not affect any other variable.
f2 turns f3 on.
f3 turns f1 on.

This structure is complete: there are no hidden variables or missing
    relationships.
The learner does not know the causal structure.
Without any information, the learner believes all 12 possible three-
    variable causal structures are equally likely a priori: f2 <- f1 ->
    f3, f1 <- f2 -> f3, f1 <- f3 -> f2, f2 -> f1 <- f3, f1 -> f2 <- f3,
    f1 -> f3 <- f2, f1 -> f2 -> f3, f1 -> f3 -> f2, f2 -> f1 -> f3, f2 ->
     f3 -> f1, f3 -> f1 -> f2, f3 -> f2 -> f1.
```

```
Importantly, causal relationships are **probabilistic**. Causes may only
    occasionally activate their effects, and variables can also turn on
    spontaneously.
Therefore, the **pedagogical intent** of your interventions becomes
    important, sometimes more so than the intended outcome.
While you cannot control the intended effect, the learner always knows
    which original variable was intervened on.
Given this, repeated interventions on the same variable are valuable.

The learner will observe two independent trials.
In each trial, all variables start off.
Then, the intervened upon variable is set to "on" (removing its incoming
    causal links), and the learner observes the (potentially noisy)
    resulting states of the other two.
The system resets between trials.

Your goal is to select two independent interventions that will be most
    informative for helping the learner identify the true causal
    structure **under the presence of noise**.
The possible intervention pairs are (select one): ["f1", "f1"], ["f1", "
    f2"], ["f1", "f3"], ["f2", "f2"], ["f2", "f3"], ["f3", "f3"].

Explain your pedagogical reasoning.
```

C PRIMITIVE CONCEPT LEARNING

C.1 TASK INTERFACE

We obtained at least 30 valid responses per unique curriculum condition. Figure 7 walks through the full participant flow for the primitive concept learning experiment. Participants are first introduced to the alien stimulus space, work through worked examples illustrating positive and negative instances, and then complete trials in which they identify the target concept from a drop-down menu given two labeled examples.

C.2 STIMULI SPACE

Table 4: Feature dimensions and values. Four body parts $\times$ three features $\times$ three values yields $27^4 = 531{,}441$ unique stimuli.

Feature	Values
Color	red, yellow, blue
Shape	circle, square, triangle
Texture	solid, striped, dotted

Stimuli are shared across the primitive and compositional tasks. Each stimulus consists of four body parts (main body, top, left, and right appendages), each defined by three features with three possible values (Table 4). In the primitive task, concepts are defined by a single feature of a single body part, yielding $9 \times 4 = 36$ hypotheses. In the compositional task, concepts are conjunctions of features across body parts, yielding $4^{12} - 1 = 16{,}777{,}215$ hypotheses which is infeasible for Bayesian pedagogical inference (Appendix D).

C.3 LLM CURRICULUM ANALYSIS

Table 5 reports the rate of Bayes-optimal pairs and contrastive strategy preference. We observe that models consistently choose examples that are Bayes-optimal ($\geq 99.5\%$), and that these choices are highly unlikely to be a result of random chance.

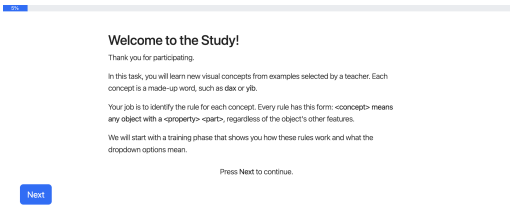
Optimal contrastive pairs hold all 11 unconstrained dimensions fixed ($n_{\text{diff}} \approx 0$), while optimal positive pairs vary all 11 ($n_{\text{diff}} = 11.00$), relative to a random baseline of $\mathbb{E}[n_{\text{diff}}] \approx 7.33$ dimensions. Under uniform random sampling of the unconstrained features, we have $P(n_{\text{diff}} = 0) = (1/3)^{11} \approx 6 \times 10^{-6}$ for contrastive pairs and $P(n_{\text{diff}} = 11) = (2/3)^{11} \approx 0.012$ for positive pairs.

Table 5: Per-model rates of Bayes-optimality, contrastive vs. positive strategy preference (Clopper–Pearson 95% CIs on contrastive rate), and mean n_{diff} (number of unconstrained feature dimensions differing between the two examples) for each strategy. The Bayesian baseline row shows the model's strategy distribution for reference.

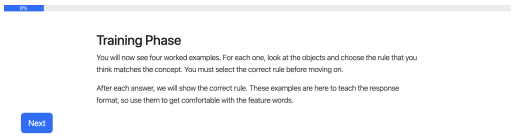
		Strategy preference		Mean n_{diff}	
Model	% optimal	% contrastive [95% CI]	% positive	Contrastive	Positive
Claude Opus 4.6	99.8	93.8 [91.6, 95.6]	6.2	0.02	11.00
Claude Sonnet 4.6	99.5	99.7 [98.8, 100.0]	0.3	0.02	11.00
GPT-5.2	100.0	98.5 [97.2, 99.3]	1.5	0.00	11.00
Gemini 3.1 Pro Preview	100.0	87.5 [84.6, 90.0]	12.5	0.00	11.00
Gemini 3 Flash Preview	100.0	99.2 [98.1, 99.7]	0.8	0.00	11.00
Bayesian baseline	—	$\approx$0.2	$\approx$99.8	—	—

C.4 PER-MODEL BEHAVIORAL RESULTS

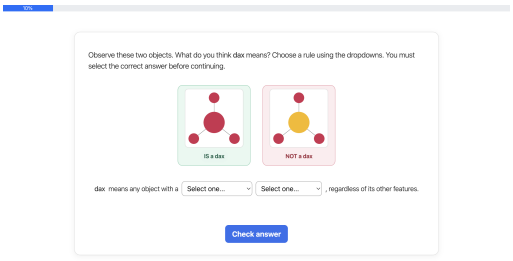
Figure 8 disaggregates the aggregate behavioral results from the main text. Three of five LLMs significantly outperform the Bayesian curriculum on accuracy (Claude Opus 4.6: $\Delta = +7.7$pp, $p =$



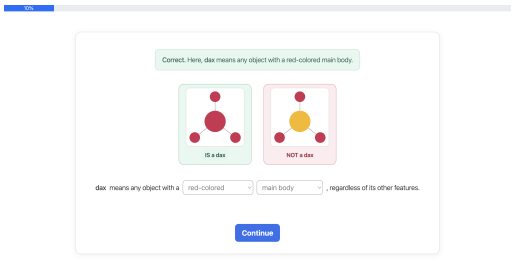
(a) Initial task instructions introducing the alien stimulus space.



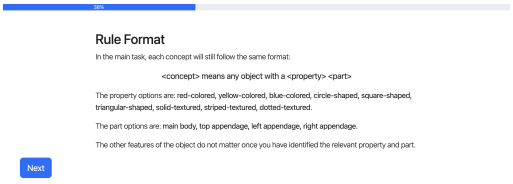
(b) Familiarization with the four body parts and three feature dimensions (color, shape, texture).



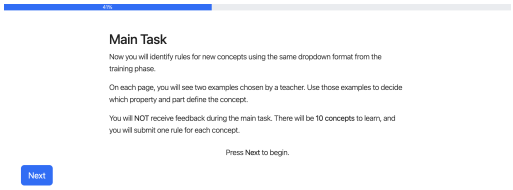
(c) Worked example illustrating a positive instance of a sample concept.



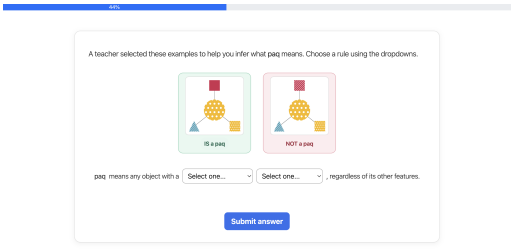
(d) Worked example illustrating a negative instance of the same sample concept.



(e) Instructions explaining that each concept is defined by a single feature on a single body part.



(f) Final pre-trial instructions describing the dropdown concept selection.



(g) Main trial interface: two labeled examples and the drop-down menu of candidate concepts.

Figure 7: Participant interface for the primitive concept learning experiment, presented in sequence: (a–b) introduction to the stimulus space, (c–d) worked positive and negative examples, (e–f) task instructions, and (g) the main trial interface.

0.001; Gemini 3.1 Pro Preview: $\Delta = +5.8$pp, $p = 0.018$; Gemini 3 Flash Preview: $\Delta = +5.4$pp, $p = 0.029$; Fisher's exact); the remaining two trend higher but do not reach significance. All five LLMs yield significantly faster response times than the Bayesian curriculum (Mann–Whitney on raw RT, all $p < 0.001$).

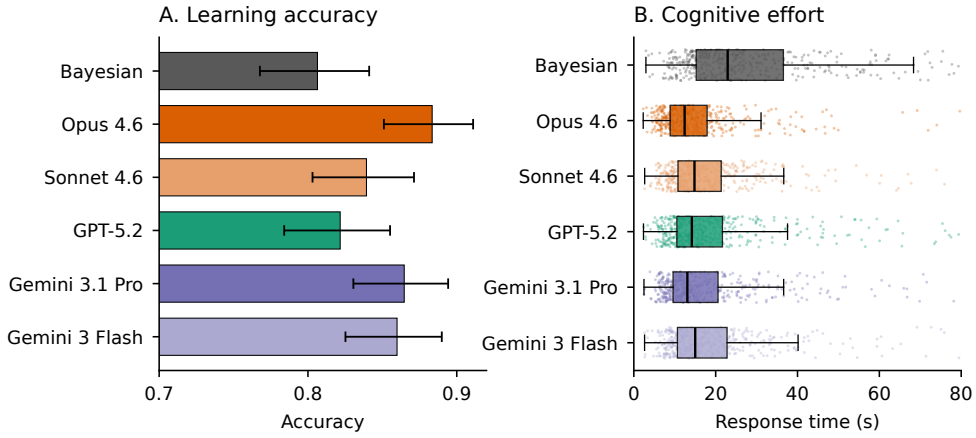


Figure 8: Per-model behavioral results. **(A)** Accuracy on target concept identification with exact binomial 95% CIs. **(B)** Response time distributions; boxes show IQR with median marked, whiskers extend to 1.5×IQR.

C.5 Prompt

```
You are an expert in concept learning. Your task is to select teaching
    examples for a single concept that will be presented to a human
    learner.

**Stimuli Space**
Stimuli in this task are visual objects. Each stimulus is composed of 4
    spatially arranged parts:
- Main body: large, displayed at center
- Top appendage: small, displayed above the main body
- Left appendage: small, displayed at bottom-left
- Right appendage: small, displayed at bottom-right

Each part independently varies along the same feature dimensions:
{
  "color": ["red", "yellow", "blue"],
  "shape": ["circle", "square", "triangle"],
  "texture": ["solid", "striped", "dotted"]
}

**Concept to Teach**
The concept "yan" is defined by the following constraints:
{
  "right appendage texture": "dotted"
}
A stimulus is an instance of "yan" if and only if it matches ALL
    constraints above. Feature dimensions not listed are unconstrained --
     any value is acceptable.

**Learner Assumptions**
The learner has no prior knowledge of this concept, but has been told
    that the concept is defined by exactly ONE feature dimension (e.g.,
    one specific part's color, shape, or texture).
```

```
Examples will be visually presented as compositional stimuli -- the
    learner sees the entire alien (all 4 parts) for each example. The
    learner has limited working memory and must infer the concept's
    defining features from the examples alone, without corrective
    feedback.

**Your Task**
Select exactly 2 labeled examples (positive or negative instances of "yan
    "). Each example must specify a value for all 12 features (color,
    shape, texture for each of the 4 parts).

Explain your pedagogical reasoning, and how your chosen examples help the
     learner identify the defining concept. Then, describe each example
    in detail -- specifying all 12 feature values and the label for each.
```

D Compositional Concept Learning

D.1 Task Interface

Figure 9 walks through the full participant flow for the compositional concept learning experiment. Participants are introduced to the rule-construction widget, complete two practice trials of increasing complexity, and then proceed to the main task in which they observe four labeled examples per step and submit a rule capturing the underlying concept.

D.2 Bayesian Intractability

In the compositional concept learning task, concepts refer to any conjunction of features localized across any subset of body parts. There are therefore $4^{12} - 1 = 16{,}777{,}215$ hypotheses (subtracting the empty set). The compositional task is intractable for Bayesian pedagogy not only because of the increased size of the hypothesis space, but also because we allow four examples to be shown instead of two, resulting in $531{,}441^4 \approx 8 \cdot 10^{22}$ possible example sets to iterate over.

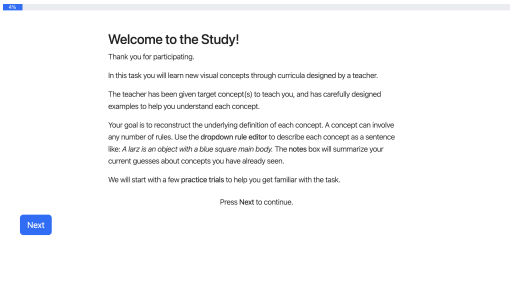
D.3 Curriculum Generation

In the compositional task, LLMs assign their own concept names as part of curriculum design. To prevent semantic leakage, we replace generated names with arbitrary nonsense words.

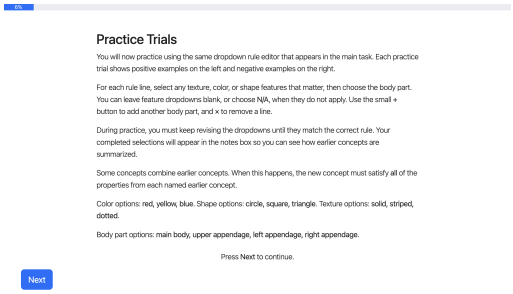
D.4 Classifying Decompositions

We use the following procedure to classify the different decompositions LLMs formulate in their curricula. For each LLM-generated curriculum, let *nested subconcepts* denote the property that at least one non-target step has a non-empty dependency list (i.e., a subconcept that itself depends on another previously taught concept), and let *target-to-target dependency* denote the property that at least one target step ($\in \{zep, tor, dem, wug\}$) has a dependency list containing another target concept. Curricula are then classified according to the following rule:

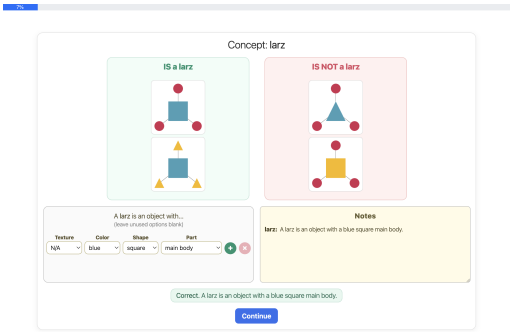
1. **Flat:** neither nested subconcepts nor target-to-target dependencies. Subconcepts are taught as standalone building blocks, and targets are independent of one another.
2. **Progressive:** target-to-target dependencies but no nested subconcepts. There is no nested subconcept hierarchy, but later targets build on earlier targets.
3. **Hierarchical:** nested subconcepts, regardless of whether target-to-target dependencies are present. This category merges two sub-cases:
   - *Tree*: nested subconcepts but no target-to-target dependencies (subconcepts organized hierarchically; targets independent of one another).
   - *Tree + progressive*: both nested subconcepts and target-to-target dependencies.



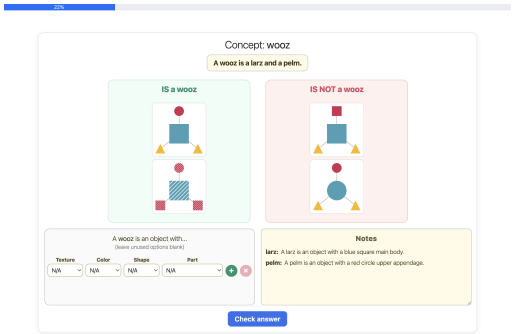
(a) Initial instructions introducing the multi-step compositional task.



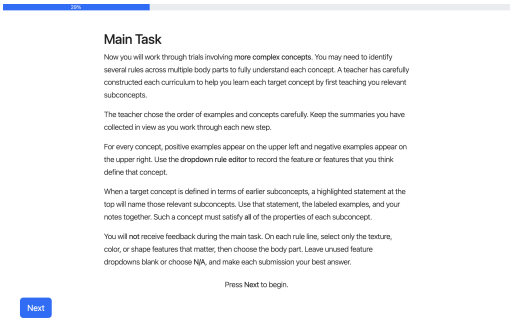
(b) Continuation of pre-practice instructions describing the rule-construction widget.



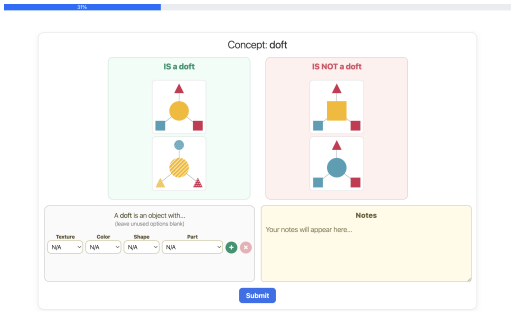
(c) First practice trial featuring a primitive (single-feature) target concept.



(d) Second practice trial featuring a compositional (multi-feature) target concept.



(e) Transition screen presenting Phase 2 instructions before the main trials.



(f) Main task trial showing four labeled examples and the rule-submission widget.

Figure 9: Participant interface for the compositional concept learning experiment: (a–b) initial instructions, (c–d) practice trials, (e) Phase 2 instructions, and (f) main task trials.

Table 6: Per-model distribution of curriculum decomposition strategies. Each cell reports the count and proportion of that model's curricula classified as Flat, Progressive, or Hierarchical (Appendix D.4). The Hierarchical column is further split into its Tree (nested subconcepts only) and Tree+Progressive (nested subconcepts and target-to-target dependencies) subcases.

				Hierarchical		
Model	n curricula	% Flat	% Progressive	% Tree	% Tree+Prog.	% Hier. total
Claude Opus 4.6	49	42.9%	57.1%	0.0%	0.0%	0.0%
Claude Sonnet 4.6	50	54.0%	40.0%	0.0%	6.0%	6.0%
GPT-5.2	50	22.0%	44.0%	18.0%	16.0%	34.0%
Gemini 3 Flash Preview	50	90.0%	6.0%	2.0%	2.0%	4.0%
Gemini 3.1 Pro Preview	50	98.0%	0.0%	2.0%	0.0%	2.0%

D.5 Per-Model Decomposition Strategies

D.6 Subconcept Spatial Localization

Table 7: Per-model statistics over all non-target subconcepts declared across the compositional experiment. We list the number of total curriculum steps and subconcepts generated by the model, as well as the number of feature constraints per subconcept. Finally, we list the percentage of subconcepts that were localized to a single spatial area ('body part'). All models had no primitive (1-feature) subconcepts.

Model	# steps	# subconcepts	mean # constraints	% single spatial area
Claude Opus 4.6	375	179	2.04	97.8%
Claude Sonnet 4.6	400	200	2.00	100.0%
GPT-5.2	393	192	2.08	95.8%
Gemini 3 Flash Preview	333	133	2.92	54.1%
Gemini 3.1 Pro Preview	402	202	2.02	99.0%

Models tend to group the visual concepts through spatial units. The minority of subconcepts that span more than a single body part are also spatially structured, binding together two whole parts at a time. Out of all models, Gemini 3 Flash Preview provides more subconcepts that span two whole parts instead of concentrating on a single spatial location.

D.7 Per-Strategy Behavioral Results

Table 8 reports per-strategy accuracy on target trials and the corresponding contrast vs. the target-only baseline. Strategy labels are assigned at the curriculum level using the decomposition rule specified in Appendix D.4. Contrasts are estimated jointly in a logistic GLMM with random intercepts for participant and concept:

$$\texttt{accuracy} \sim \texttt{strategy} + (1 \mid \texttt{participant}) + (1 \mid \texttt{concept}),$$

fit on the pre-registered clean cohort ($n = 1{,}508$ target trials from 377 participants).

Table 8: Target-trial accuracy by strategy and contrast vs. target-only baseline. "Target-only" is the reference level. $\hat{\beta}$, OR, z, and p are from a single logistic GLMM with random participant and concept intercepts.

Strategy	Accuracy	n trials	$\hat{\beta}$ (logit)	OR	95% CI	p
Target-only	36.5%	688	—	—	—	—
Flat	52.8%	536	+1.69	5.43	[2.20, 13.45]	< 0.001
Progressive	44.0%	216	+0.78	2.19	[0.68, 7.08]	0.190
Hierarchical	48.5%	68	+1.19	3.28	[0.51, 21.02]	0.209

Flat decomposition produces the only contrast that survives at $\alpha = 0.05$. Progressive and hierarchical decomposition trend in the same direction with smaller effect sizes; their wide confidence intervals reflect both the smaller cell sizes and substantial between-participant variance ($\sigma^2_{\text{participant}} = 10.0$ on the logit scale).

F1-score contrast. The parallel feature-level F1 contrast (decomposed vs. target-only baseline, pooled across strategies) is fit as

$$\texttt{F1} \sim \texttt{condition_c} + (1 \mid \texttt{participant}) + (1 + \texttt{condition_c} \mid \texttt{concept}),$$

yielding $\hat{\beta} = 0.074$, 95% CI $[0.024, 0.125]$, Satterthwaite $t(218) = 2.88$, $p = 0.0043$.

D.8 Practice Effects and Subconcept-Target Performance Correlation

Table 9: Per-step accuracy in the decomposed (subconcept trials) and target-only (target trials) conditions. "Accuracy" is the proportion of trials on which the participant's submitted rule exactly matches the ground-truth concept definition (Wilson 95% CIs). "F1" is the mean feature-level F1 between the submitted and ground-truth rule sets.

Condition / step	n trials	Accuracy	95% CI	F1
Decomposed condition, subconcept trials				
Step 1	206	74.8%	[68.4, 80.2]	87.5%
Step 2	199	72.9%	[66.3, 78.6]	81.8%
Step 3	185	69.7%	[62.8, 75.9]	79.0%
Step 4	173	72.3%	[65.2, 78.4]	79.7%
Target-only baseline, target trials				
Step 1	174	29.9%	[23.6, 37.1]	68.3%
Step 2	173	37.0%	[30.2, 44.4]	72.4%
Step 3	173	38.7%	[31.8, 46.2]	73.0%
Step 4	172	39.5%	[32.5, 47.0]	73.8%

Subconcept-target coupling. For each participant in the decomposed condition with at least one subconcept trial and one target trial, we compute the per-participant mean accuracy on subconcept trials and on target trials. The Spearman rank correlation between these two per-participant scores is $\rho = 0.598$, $p < 0.001$ ($n = 203$ participants; bootstrap 95% CI $[0.50, 0.69]$, $B = 5{,}000$ resamples). The same analysis using feature-level F1 instead of exact-match accuracy yields a similar correlation ($\rho = 0.64$, $p < 0.001$). Participants who performed well on the subconcepts also tended to perform well on the corresponding compositional targets, consistent with subconcept understanding contributing to target understanding.

D.9 Prompt

The curriculum scaffolding prompt is included below. Prompts to generate examples for subconcepts and targets based on this scaffolding follow the exact same form as the prompt in Appendix C.5, but specify providing four examples instead of two; this same prompt is also used to generate data for the baseline condition, which relies on example selection alone to build understanding of targets.

```
You are an expert in concept learning and curriculum design. Your task is
     to design the structure of a teaching curriculum that will help a
    learner master a set of novel concepts. A separate system will
    generate teaching examples for each step -- your responsibility is
    only to determine which concepts to teach, in what order, and how to
    describe them to the learner.

**Stimuli Space**
[identical to primitive task prompt; see Appendix~\ref{app:primitive-
    prompt}]

**Concept Space**
```

```
The learner must ultimately master the following 4 target concepts. Each
    concept is defined as a set of feature constraints over the stimulus
    space. A stimulus is an instance of the concept if and only if it
    matches ALL of the feature constraints listed. Feature dimensions not
     listed are unconstrained -- any value is acceptable.
{
  "zep": {"main body color": "yellow", "main body shape": "circle", "top
    appendage color": "red", "top appendage shape": "triangle", "left
    appendage color": "red", "left appendage shape": "circle"},
  "tor": {"main body color": "yellow", "main body shape": "circle", "top
    appendage color": "red", "top appendage shape": "triangle", "right
    appendage color": "blue", "right appendage shape": "square"},
  "dem": {"main body color": "yellow", "main body shape": "circle", "left
     appendage color": "red", "left appendage shape": "circle", "right
    appendage color": "blue", "right appendage shape": "square"},
  "wug": {"top appendage color": "red", "top appendage shape": "triangle
    ", "left appendage color": "red", "left appendage shape": "circle", "
    right appendage color": "blue", "right appendage shape": "square"}
}

**Learner Assumptions**
The learner has no prior knowledge of the concepts or the structure of
    the concept space. The learner retains knowledge from earlier steps
    and can build on it when learning subsequent concepts. The learner
    has limited working memory and will learn each concept from a small
    number of examples.

**Curriculum Design**
Design an ordered teaching curriculum that will maximize the learner's
    ability to master all 4 target concepts. For each step, provide:
1. A CONCEPT NAME -- the target concept name if applicable, otherwise an
    arbitrary nonsense word.
2. A CONCEPT DEFINITION -- the set of feature constraints defining this
    concept.
3. RELEVANT PRIOR CONCEPTS -- concept names from previous steps relevant
    to this one (empty list if none).

You may introduce additional scaffolding steps beyond the four targets.
    Every target concept must appear as a step.

Explain your pedagogical reasoning for the curriculum structure.
```

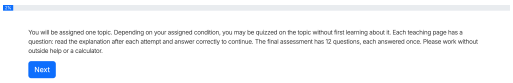
E NATURALISTIC CONCEPT LEARNING

E.1 TASK INTERFACE

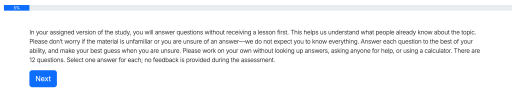
Figure 10 walks through the full participant flow for the naturalistic concept learning experiments. All participants first read the study instructions; participants in the no-curriculum condition then proceed directly to the assessment. In the curriculum conditions, participants work through a sequence of teaching pages, each ending with a multiple-choice practice question that must be answered correctly to continue, with an explanation shown after each attempt. Participants then complete a twelve-question multiple-choice assessment without feedback, describe their learning strategy, and see their final score. Figure 11 shows example teaching pages from the three curriculum conditions of Experiment 2, and Figure 12 shows example decomposed curricula from each model in Experiment 1.

E.2 CURRICULUM STRUCTURE ANALYSIS

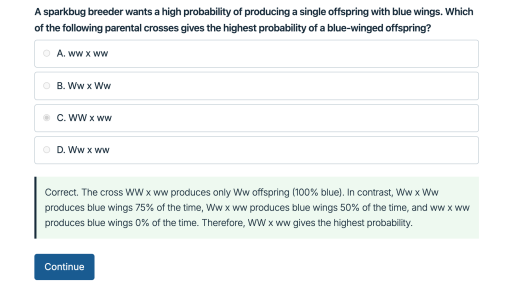
Method. For each model and domain, we generated $n = 20$ single-pass curricula (180 total). Each single-pass curriculum was scored against one of the model's four-module decomposition plans from the two-stage procedure (regenerated with the same procedure as the experimental curricula),



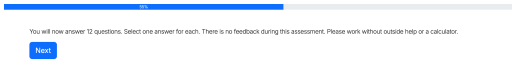
(a) Study instructions shown to all participants.



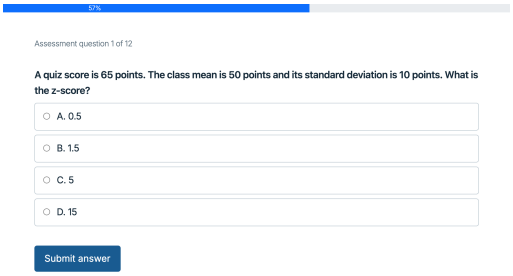
(b) Additional instructions in the no-curriculum condition, which proceeds directly to the assessment.



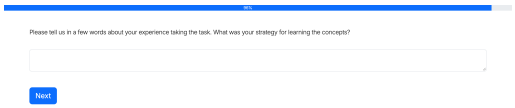
(c) Practice question at the end of a teaching page, with the explanation shown after an attempt.



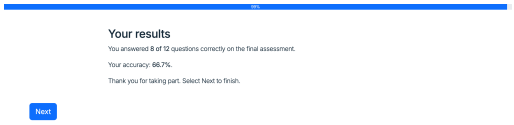
(d) Transition screen presenting the assessment instructions.



(e) Assessment question (z-scores), answered once without feedback.

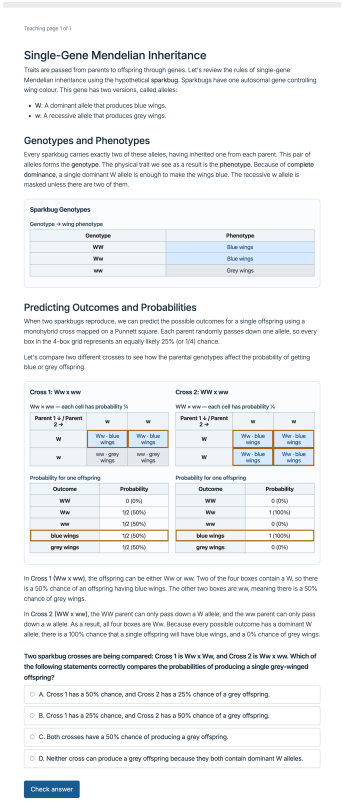


(f) Free-response question about the participant's learning strategy.



(g) Final screen reporting the participant's assessment score.

Figure 10: Participant interface for the naturalistic concept learning experiments: (a–b) instructions, (c) practice question on a teaching page, (d–e) assessment, and (f–g) post-task screens.



Teaching page 1 of 1

Single-Gene Mendelian Inheritance

Traits are passed from parents to offspring through genes. Let's review the rules of single-gene Mendelian inheritance using the hypothetical **sparkbug**. Sparkbugs have one autosomal gene controlling wing colour. This gene has two versions, called alleles:

- **W**: A dominant allele that produces blue wings.
- **w**: A recessive allele that produces grey wings.

Genotypes and Phenotypes

Every sparkbug carries exactly two of these alleles, having inherited one from each parent. This pair of alleles forms the **genotype**. The physical trait we see as a result is the **phenotype**. Because of **complete dominance**, a single dominant W allele is enough to make the wings blue. The recessive w allele is masked unless there are two of them.

Sparkbug Genotypes

Genotype → wing phenotype

Genotype	Phenotype
WW	Blue wings
Ww	Blue wings
ww	Grey wings

Predicting Outcomes and Probabilities

When two sparkbugs reproduce, we can predict the possible outcomes for a single offspring using a monohybrid cross mapped on a Punnett square. Each parent randomly passes down one allele, so every box in the 4-box grid represents an equally likely 25% (or 1/4) chance.

Let's compare two different crosses to see how the parental genotypes affect the probability of getting blue or grey offspring.

Cross 1: Ww x ww

Ww × ww — each cell has probability ¼

Parent 1 ↓ / Parent 2 →	w	w
W	Ww - blue wings	Ww - blue wings
w	ww - grey wings	ww - grey wings

Probability for one offspring

Outcome	Probability
WW	0 (0%)
Ww	1/2 (50%)
ww	1/2 (50%)
blue wings	1/2 (50%)
grey wings	1/2 (50%)

Cross 2: WW x ww

WW × ww — each cell has probability ¼

Parent 1 ↓ / Parent 2 →	w	w
W	Ww - blue wings	Ww - blue wings
W	Ww - blue wings	Ww - blue wings

Probability for one offspring

Outcome	Probability
WW	0 (0%)
Ww	1 (100%)
ww	0 (0%)
blue wings	1 (100%)
grey wings	0 (0%)

In **Cross 1 (Ww x ww)**, the offspring can be either Ww or ww. Two of the four boxes contain a W, so there is a 50% chance of an offspring having blue wings. The other two boxes are ww, meaning there is a 50% chance of grey wings.

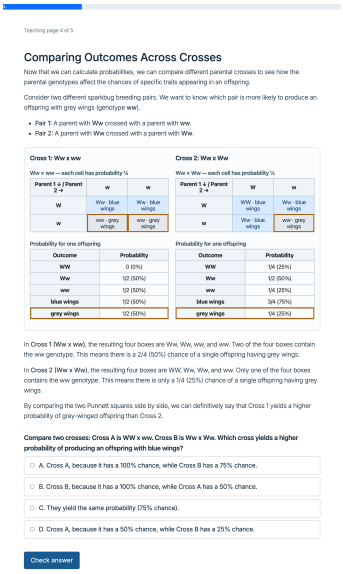
In **Cross 2 (WW x ww)**, the WW parent can only pass down a W allele, and the ww parent can only pass down a w allele. As a result, all four boxes are Ww. Because every possible outcome has a dominant W allele, there is a 100% chance that a single offspring will have blue wings, and a 0% chance of grey wings.

Two sparkbug crosses are being compared: Cross 1 is Ww x Ww, and Cross 2 is Ww x ww. Which of the following statements correctly compares the probabilities of producing a single grey-winged offspring?

- A. Cross 1 has a 50% chance, and Cross 2 has a 25% chance of a grey offspring.
- B. Cross 1 has a 25% chance, and Cross 2 has a 50% chance of a grey offspring.
- C. Both crosses have a 50% chance of producing a grey offspring.
- D. Neither cross can produce a grey offspring because they both contain dominant W alleles.

Check answer

(a) Target-only: the final, integrative module of the decomposed curriculum, presented as a single page.



Teaching page 4 of 5

Comparing Outcomes Across Crosses

Now that we can calculate probabilities, we can compare different parental crosses to see how the parental genotypes affect the chances of specific traits appearing in an offspring.

Consider two different sparkbug breeding pairs. We want to know which pair is more likely to produce an offspring with grey wings (genotype **ww**).

- **Pair 1**: A parent with **Ww** crossed with a parent with **ww**.
- **Pair 2**: A parent with **Ww** crossed with a parent with **Ww**.

Cross 1: Ww x ww

Ww × ww — each cell has probability ¼

Parent 1 ↓ / Parent 2 →	w	w
W	Ww - blue wings	Ww - blue wings
w	ww - grey wings	ww - grey wings

Probability for one offspring

Outcome	Probability
WW	0 (0%)
Ww	1/2 (50%)
ww	1/2 (50%)
blue wings	1/2 (50%)
grey wings	1/2 (50%)

Cross 2: Ww x Ww

Ww × Ww — each cell has probability ¼

Parent 1 ↓ / Parent 2 →	W	w
W	WW - blue wings	Ww - blue wings
w	Ww - blue wings	ww - grey wings

Probability for one offspring

Outcome	Probability
WW	1/4 (25%)
Ww	1/2 (50%)
ww	1/4 (25%)
blue wings	3/4 (75%)
grey wings	1/4 (25%)

In **Cross 1 (Ww x ww)**, the resulting four boxes are Ww, Ww, ww, and ww. Two of the four boxes contain the ww genotype. This means there is a 2/4 (50%) chance of a single offspring having grey wings.

In **Cross 2 (Ww x Ww)**, the resulting four boxes are WW, Ww, Ww, and ww. Only one of the four boxes contains the ww genotype. This means there is only a 1/4 (25%) chance of a single offspring having grey wings.

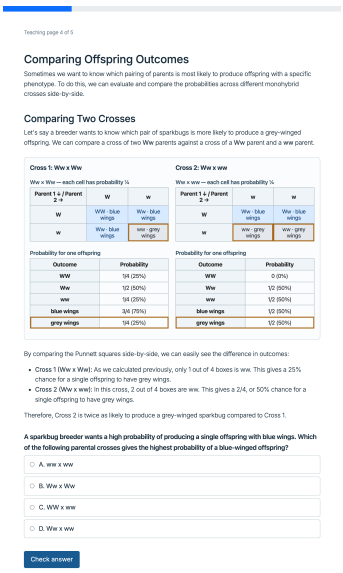
By comparing the two Punnett squares side by side, we can definitively say that Cross 1 yields a higher probability of grey-winged offspring than Cross 2.

Compare two crosses: Cross A is WW x ww. Cross B is Ww x Ww. Which cross yields a higher probability of producing an offspring with blue wings?

- A. Cross A, because it has a 100% chance, while Cross B has a 75% chance.
- B. Cross B, because it has a 100% chance, while Cross A has a 50% chance.
- C. They yield the same probability (75% chance).
- D. Cross A, because it has a 50% chance, while Cross B has a 25% chance.

Check answer

(b) Single-pass: page 4 of 5 from a curriculum generated in a single call.



Teaching page 4 of 5

Comparing Offspring Outcomes

Sometimes we want to know which pairing of parents is most likely to produce offspring with a specific phenotype. To do this, we can evaluate and compare the probabilities across different monohybrid crosses side-by-side.

Comparing Two Crosses

Let's say a breeder wants to know which pair of sparkbugs is more likely to produce a grey-winged offspring. We can compare a cross of two **Ww** parents against a cross of a **Ww** parent and a **ww** parent.

Cross 1: Ww x Ww

Ww × Ww — each cell has probability ¼

Parent 1 ↓ / Parent 2 →	W	w
W	WW - blue wings	Ww - blue wings
w	Ww - blue wings	ww - grey wings

Probability for one offspring

Outcome	Probability
WW	1/4 (25%)
Ww	1/2 (50%)
ww	1/4 (25%)
blue wings	3/4 (75%)
grey wings	1/4 (25%)

Cross 2: Ww x ww

Ww × ww — each cell has probability ¼

Parent 1 ↓ / Parent 2 →	w	w
W	Ww - blue wings	Ww - blue wings
w	ww - grey wings	ww - grey wings

Probability for one offspring

Outcome	Probability
WW	0 (0%)
Ww	1/2 (50%)
ww	1/2 (50%)
blue wings	1/2 (50%)
grey wings	1/2 (50%)

By comparing the Punnett squares side-by-side, we can easily see the difference in outcomes:

- **Cross 1 (Ww x Ww)**: As we calculated previously, only 1 out of 4 boxes is ww. This gives a 25% chance for a single offspring to have grey wings.
- **Cross 2 (Ww x ww)**: In this cross, 2 out of 4 boxes are ww. This gives a 2/4, or 50% chance for a single offspring to have grey wings.

Therefore, Cross 2 is twice as likely to produce a grey-winged sparkbug compared to Cross 1.

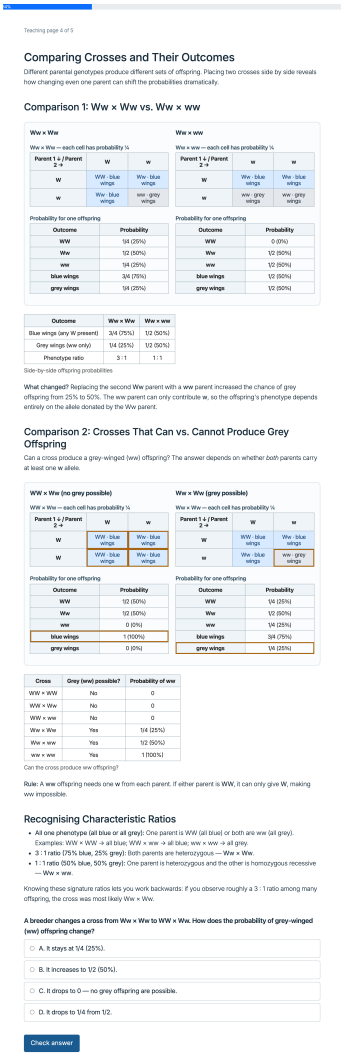
A sparkbug breeder wants a high probability of producing a single offspring with blue wings. Which of the following parental crosses gives the highest probability of a blue-winged offspring?

- A. ww x ww
- B. Ww x Ww
- C. WW x ww
- D. Ww x ww

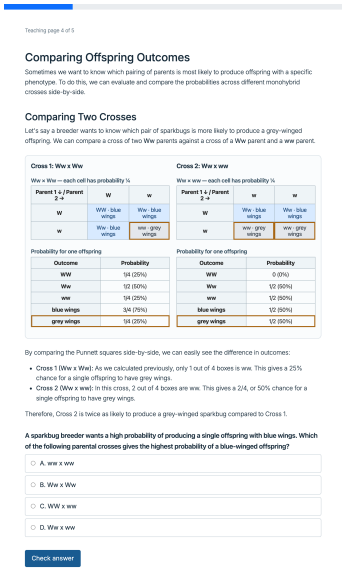
Check answer

(c) Two-stage: page 4 of 5 from a curriculum generated module by module.

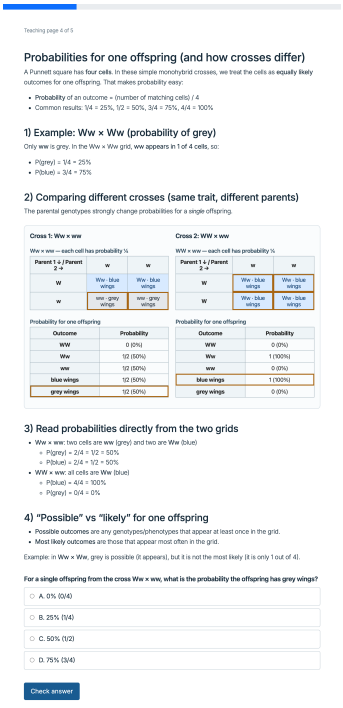
Figure 11: Example teaching pages from the three curriculum conditions of Experiment 2 (Mendelian genetics; all curricula generated by Gemini 3.1 Pro Preview). Each page ends with a practice question that must be answered correctly to continue.



(a) Claude Opus 4.6.



(b) Gemini 3.1 Pro Preview.



(c) GPT-5.2.

Figure 12: Example teaching pages from two-stage curricula generated by each model in Experiment 1 (Mendelian genetics).

in which each module lists the modules it depends on. An LLM evaluator (Gemini 3.1 Flash-Lite) segmented each curriculum into instructional units (semantically coherent concepts, skills, or integrated problems) without access to the plan, then aligned the units to the plan's topics. Annotation failed for 11 curricula, which are excluded.

Table 10: Prerequisite order in single-pass curricula relative to each model's two-stage decomposition plans: % of prerequisite relations whose prerequisite module was first taught in an earlier unit than its dependent, in the same unit, or later (Rev.). Chance: expected "earlier" under random unit order; p: Wilcoxon signed-rank test of Earlier against Chance across curricula. Obj.: objective pairs introduced in the order listed in the prompt.

Model	n	Earlier	Same	Rev.	Chance	p	Obj.
Claude Opus 4.6	53	94.2	5.5	0.3	47.3	<0.001	93.3
Gemini 3.1 Pro	60	85.4	13.7	0.9	43.1	<0.001	94.4
GPT-5.2	56	70.8	23.2	6.0	38.4	<0.001	90.6
All	169	83.3	14.3	2.4	42.9	<0.001	92.8

E.3 Behavioral Results

Table 11: Assessment accuracy (%) by domain, condition, and the LLM used to generate the curricula (Experiment 1). Cells report the mean of participant-level accuracies, with SEM in parentheses. Chance is 25%.

			Two-stage, by LLM		
Domain	No curriculum	Two-stage (all)	Opus 4.6	GPT-5.2	Gemini 3.1 Pro
Logic circuits	28.1 (5.9)	82.6 (3.3)	79.8 (5.9)	83.3 (6.7)	84.8 (5.0)
Mendelian genetics	67.2 (7.2)	77.4 (3.4)	76.6 (6.3)	76.7 (4.8)	79.2 (7.1)
z-scores	56.0 (8.0)	78.9 (4.1)	82.1 (7.2)	80.7 (7.3)	73.8 (6.8)
Overall	50.4 (4.7)	79.6 (2.1)	79.5 (3.7)	79.9 (3.6)	79.4 (3.6)
n participants	48	136	48	43	45
Median RT (s)	15.9	17.1	18.1	15.4	17.1

Table 12: Assessment accuracy (%) by domain and curriculum condition (Experiment 2). All curricula were generated by Gemini 3.1 Pro Preview. The no-curriculum group is shared with Experiment 1. Cells report the mean of participant-level accuracies, with SEM in parentheses. Chance is 25%.

Domain	No curriculum	Target-only	Single-pass	Two-stage
Logic circuits	28.1 (5.9)	64.7 (4.8)	79.6 (4.7)	78.2 (4.6)
Mendelian genetics	67.2 (7.2)	73.1 (4.1)	79.6 (3.9)	77.1 (4.6)
z-scores	56.0 (8.0)	74.7 (4.8)	80.5 (5.4)	83.9 (4.1)
Overall	50.4 (4.7)	70.8 (2.7)	79.9 (2.7)	79.7 (2.6)
n participants	48	89	88	89
Median RT (s)	15.9	19.0	17.8	17.2

Pooled models and preregistered GLMMs. Following our preregistration for the first experiment (no curriculum vs. two-stage curricula from Claude Opus 4.6, Gemini 3.1 Pro Preview, and GPT-5.2), the pooled model fit to $n = 2{,}100$ trials from 184 participants (excluding 108 trials under the preregistered exclusion criteria) is

$$\texttt{Accuracy} \sim \texttt{Condition} + (1 \mid \texttt{LLM}) + (1 \mid \texttt{Domain}),$$

where `Condition` is the two-stage vs. no-curriculum contrast (two-stage $= +0.5$, no curriculum $= -0.5$). Because the no-curriculum group has no generating model, the LLM intercept applies to

curriculum trials only. For the second experiment (curriculum ablations, with all curricula generated by Gemini 3.1 Pro Preview), the pooled model fit to $n = 3{,}617$ trials from 314 participants (excluding 163 trials under the preregistered exclusion criteria) is

```
Accuracy ~ Condition + (1 | Domain),
```

with successive-difference contrasts over the no-curriculum, target-only, single-pass, and two-stage conditions, in that order. The no-curriculum group is shared with the first experiment.

Table 13: Preregistered mixed-effects logistic regressions (log-odds scale). Random-intercept SDs: Experiment 1 (LLM: 0.00, Domain: 0.12); Experiment 2 (Domain: 0.17). †One-sided, as preregistered; other p-values are two-sided.

Experiment	Effect	$\hat{\beta}$	SE	95% CI	z	p
Experiment 1	Two-stage − No curriculum	+1.31	0.11	[1.10, 1.52]	12.14	<0.001
Experiment 2	C1: Target-only − No curriculum	+0.79	0.11	[0.58, 1.01]	7.14	<0.001†
	C2: Single-pass − Target-only	+0.53	0.11	[0.32, 0.74]	5.03	<0.001†
	C3: Two-stage − Single-pass	−0.06	0.11	[−0.27, 0.16]	−0.51	0.609

E.4 PROMPT

We show a sample prompt for logic circuits; Mendelian genetics and z-scores follow suit. The objective is the target concept or skill for the student to master, and the constraints limit the scope of the curriculum. We further specify setting in order to standardize the curricula.

```
Subject: Logic circuits with AND, OR, and NOT gates
Setting: All examples use a small control panel whose binary switches
    are named **A**, **B**, and **C**, where 1 means a switch is on and
    0 means off. Gates combine those signals; name any intermediate wire
    **P** or **Q**, and call the final output **OUT**.
Objectives:
- gates: Interpret binary inputs and AND, OR, and NOT gate behavior.
- tracing: Trace outputs through up to three connected gates.
- missing_input: Infer a missing input when the information determines
    it uniquely.
- comparison: Compare the outputs of small circuits for stated input
    settings.
Constraints:
- Assume basic arithmetic but no prior logic or electronics knowledge.
- Use only binary inputs 0 and 1, two-input AND and OR gates, and
    one-input NOT gates.
- Describe each circuit in words or as a short list of connections; name
    every input and state what feeds each gate and where its output goes.
- Keep circuits to at most three gates and at most three inputs, so any
    output can be traced mentally.
- Exclude XOR, NAND, NOR, and XNOR gates, Boolean algebra notation and
    its laws, De Morgan's laws, truth-table minimisation, timing,
    clocks, memory, and how a gate is built physically.
```

Single-pass instructions. Used for the single-pass curricula in the second experiment.

```
Teach a complete novice this subject through exactly five pages: four
    teaching pages followed by one final synthesis. Cover every
    objective with clear rules and fully worked examples using the fixed
    setting and notation above.

The final synthesis must stand on its own. Explain the notation and
    rules it needs and show representative worked examples supporting
    all the objectives. Do not refer to earlier pages, assume the
    learner has read them, or discuss the preceding curriculum. Present
    the synthesis as a self-contained explanation of the subject. Write
    every page for the learner, not about your writing process.
```

Two-stage instructions. In the first stage, the model produced a decomposition plan: four modules, each with a title, description, topics, and the modules it depends on. In the second stage, the model received the single-pass prompt above with the plan appended:

```
Use this concept plan to determine what to teach and its order:
<decomposition plan from the first stage, as JSON>
```

Freeform single-pass instructions. Used for the additional single-pass curricula in the curriculum structure analysis (Appendix E.2), replacing the single-pass instructions above.

```
Teach a complete novice this subject so they can achieve all the
    objectives. Choose the organization, sequence, and amount of
    explanation you consider appropriate. Use clear rules and fully
    worked examples with the fixed setting and notation above. You may
    divide the lesson into pages wherever useful; there is no required
    number of pages. Choose whether a concluding summary would be
    useful. Write for the learner, not about your writing process.
```