

Reasoning Models of Artificial Intelligence Think Like Humans Do

Jiawei Xiang

jiawei.xiang@cgu.edu

Claremont Graduate University

Elbert Blakely

Florida Institute of Technology

Henry D. Schlinger, Jr.

California State University Los Angeles

Perspective

Keywords: mediational theory, derived relations, artificial intelligence, large language models, reasoning models

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11132673/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

A mediational theory of behavioral relations offers a parsimonious account of performances commonly described as derived or emergent following matching-to-sample (MTS) procedures and verbal instructions. It follows logically then, that if “derived” is taken to mean untrained or unreinforced, then no such relations exist. Apparent stimulus equivalence and derived relational responding can be explained as directly, though not explicitly, reinforced mediating behaviors, including overt or covert echoic and intraverbal responses maintained by social or automatic reinforcement (e.g., parity). The distinction between explicit vs. direct training emphasizes that relations not explicitly targeted by researchers may nevertheless be reinforced and thus directly established. Empirical research with humans and nonhumans, as well as analyses of mediated generalization and joint control, supports this interpretation. Recent developments in artificial intelligence show that large language models now engage in behaviors that resemble the behaviors described by a mediational theory of behavioral relations. Thus, computer programmers have arrived at the same conclusion that behavior scientists have with respect to mediating behaviors. We argue that that is no coincidence because that is how problems are most effectively solved, whether by human or machine. In the present article we 1) summarize a mediational theory of behavioral relations; 2) describe recent developments in artificial intelligence, particularly the evolution of reasoning large language models that employ chain-of-thought processing and reinforcement learning; and 3) point out that these systems demonstrate that stepwise, reinforced mediation enhances problem solving and therefore mirror core aspects of a mediational theory of behavior.

A Brief Overview of a Mediational Theory

In a recent article, two of us (Schlinger & Blakely) described how a mediational theory parsimoniously explains what are often referred to as *derived* or *emergent* relations that result from matching-to-sample (MTS) procedures with humans and nonhumans (Schlinger & Blakely, 2024), often called stimulus equivalence (e.g., Green & Saunders, 1998; Sidman & Tailby, 1982) or derived relational responding (DRR) (e.g., Hayes et al., 2001). We concluded that if by *derived* or *emergent*, one means untrained or unreinforced, then there is no such thing as a derived or emergent relation. We argued that all such relations are indeed directly reinforced even though the relations are not necessarily targeted for reinforcement by the researcher.

Consider a simple example. Suppose a parent says to a reasonably verbal child, “Point to the car” and reinforces the correct pointing. As a result, the child can now say “car” when asked “What is that?” and can also select a picture of a car from an array of pictures of other objects when asked to. We and others (e.g., Horne & Lowe, 1996; Miguel, 2016) have argued that when the child hears “car” she echoes it and that response form (“car”) is reinforced at the same time as the pointing response, thus converting the echoic response form into a tact. The child’s response “car” to a car or to a picture of a car was not derived nor did it emerge; it was directly reinforced even though the relation was not targeted by the researcher for reinforcement. We also argued that the reinforcement for the echoic form of the response

need not necessarily come from the parent; automatic reinforcement from the match between what the child heard and what the child hears herself say—that is, parity (Palmer, 1996)—is sufficient to establish the tact.

More recently, we extended a mediational theory of behavioral relations resulting from MTS procedures to those resulting from verbal stimuli (i.e., instructions or rules) (Schlinger & Blakely, 2026). Our argument was the same as it was with respect to relations that result from MTS procedures: that what seem to be emergent or derived relations are, in fact, directly, though not explicitly, reinforced. For example, suppose an elementary school teacher shows a picture of a sloth to the students and says, “This is a sloth.” Without any further instruction, and if the children were listening (see Schlinger, 2008), the children can now say “sloth” when asked by the teacher “What is that called?” They can also point to the picture of a sloth in an array of other animals when asked to. Our argument for how this happens is the same as our argument for how MTS procedures appear to produce derived relations: the children echo “sloth” out loud when they hear the teacher say it, and that response form is reinforced directly by the teacher in the presence of the picture. If the echoic is sub-audible, there is automatic reinforcement from the match between what the children hear from the teacher and what they “hear” from themselves. Once again, the resulting relations are not derived nor do they emerge, if by those terms one means not reinforced.

The central claim of both articles, then, is that mediating behavior, including echoic and intraverbal responses that occur during instruction and in response to important events in daily life, is directly reinforced as part of the contingencies in effect, even when that reinforcement is not explicitly arranged by the instructor or researcher.

In both our articles, we distinguished between direct and explicit training. However, this distinction is ambiguous. So, here we try to clarify the distinction between relations that are targeted by the researcher or teacher for reinforcement vs. those that are not targeted for reinforcement. In relations that are targeted by the researcher or teacher, the reinforcement is programmed, and therefore explicit. In relations that are not targeted by the research or teacher, there is no programmed reinforcement. In such cases, non-targeted relations are incidentally reinforced. In both cases, however, the reinforcement is direct; that is, relations are strengthened as a result. Direct training can include programmed reinforcement when, for example, a clinician arranges a reinforcer for mands. Direct training, however, is often not programmed. Responses that are not targeted by researchers or clinicians might nevertheless be directly reinforced, therefore creating new relations. These relations are akin to what Sidman (2000) referred to as undefined. As he wrote, “the subject may do other things (rspX) instead of the defined response. The subject’s responses may produce undefined consequences (rfX); also, undefined discriminative and conditional stimuli (XX) may control what the subject does” (p. 129). The point of our previous articles was that in experiments that seemed to show emergent relations (i.e., equivalence, naming, or DRR), such relations were, in fact, directly trained; that is, the resulting relations were undefined or not explicitly targeted by the researcher or practitioner but still resulted from reinforcement.

We have so far briefly summarized a mediational theory to explain results from MTS procedures that are said to establish stimulus equivalence, as well as results of studies involving verbal stimuli. In addition to making the case experimentally and theoretically for a mediational theory, we also noted implications of such a theory for clinical practice. In particular, instead of delivering sometimes hundreds of multiple-exemplar trials, practitioners could simply identify possible mediating behaviors and target those for training, thus circumventing lengthy procedures (e.g., Zhou et al., 2025). In the present report, however, we describe a different sort of implication, namely that current large language models (LLMs) of artificial intelligence (AI) mirror the behaviors suggested by a mediational theory of problem solving.

Reasoning Models of Artificial Intelligence

The logic of verbal mediation, as we have described it, is straightforward: complex behavioral performances that appear to emerge without reinforcement are, in fact, the product of intermediate—mediating—responses that were directly, though not explicitly, reinforced. This account offers far-reaching implications for problem-solving, including performance in stimulus equivalence tasks and reasoning. Moreover, and perhaps most importantly, a mediational theory involves a set of specific skills that when taught, allows learners to effectively respond to a variety of problems.

Modern versions of AI also solve complex problems and exhibit “reasoning” capabilities. Individuals who are actively involved in programming LLMs to accomplish these tasks have arrived at a fascinating conclusion: the most effective way to program a computer to solve a problem or engage in some sort of task that requires reasoning is to use a sequence of mediating steps that are strikingly similar to those described in our previous articles (Schlinger & Blakely, 2024, 2026) and summarized in this paper.

A Brief History of Reasoning Models of AI

The idea that a machine might solve complex problems by working through steps like humans do did not arrive all at once. The earliest learning machines already operated on a behavioral principle, which is learning with feedback. Rosenblatt's (1958) *perceptron*, an early artificial system, learned to associate inputs with output by strengthening the connections that produced correct outputs and weakening those that produced errors, a procedure any behavior analyst will recognize as selection by consequences (Skinner, 1981). However, this approach could handle only simple discriminations. Another step in the development of more effective machines was *backpropagation* (Rumelhart et al., 1986). In this method, when an output has an error, that error is passed backward through the network, and the strength of every connection that contributed to it is adjusted so as to reduce the error in the future.

For decades such networks remained small, less because the principle was inadequate than because the training that could be arranged was too brief. The real constraint was computing power, and it eased once researchers began training networks on the graphics processors originally built for video games (Krizhevsky et al., 2017); performance then improved in an orderly way as models, data, and computation grew (Kaplan et al., 2020). The *transformer architecture* (Vaswani et al., 2017), a design that read a whole

passage at once and learned which earlier words matter most for predicting the next, made it practical to train on enormous amounts of text, and by 2020 models such as GPT-3 could perform many verbal tasks from only a few examples (Brown et al., 2020). Even so, these systems still answered in a single step, predicting the most probable continuation of the text with no intermediate behavior of their own. Because their knowledge resided in the strengths of billions of internal connections rather than in rules a programmer had written, even their designers could not say why they responded as they did, the property that came to be called the *black box* (Castelvecchi, 2016). The turn toward step-by-step processing came not from theory but from trial and error: researchers found that prompting a model to show its work, or simply instructing it to think step by step, sharply improved performance on multistep problems (Kojima et al., 2022; Wei et al., 2022). In behavioral terms, the field had discovered what mediational theorists have maintained, that intermediate verbal responses make the solution of complex problems more probable, and, as we describe below, that these mediating steps become far more effective when reinforcement is applied to each step rather than to the final answer alone.

The Evolution of Reasoning LLMs

Problem-solving with humans does not occur in a vacuum; a history of reinforcement that involves a variety of skills is required. Imagine the range of foundational skills that are required to solve for x in the Eq. $3x + 9 = 24$, or to respond to a scenario such as “You are driving 60 mph, so how long will it take to get to Cleveland, which is 200 miles from here?” Just as humans require foundational knowledge to solve problems, so do computers. Below, we explore the recent development of AI models, specifically, the ways in which that foundational knowledge has been built.

To build foundational knowledge, a model is first exposed to an enormous sample of text, from which it learns the statistical structure of language but nothing about which of its outputs are useful. It is then given examples of preferred responses, which narrow that repertoire toward answers people find helpful and safe. Only at the third stage do the model's own outputs begin to produce consequences that select them, first through the judgments of human raters and then through automatic verification of whether an answer, or each step leading to it, is correct. Each stage supplies something the preceding one did not, and it is the last of these that brings the training of these systems into contact with a mediational account.

Unsupervised Pretraining

The first stage of building foundational knowledge is called *unsupervised (or self-supervised) pretraining*. Information from a variety of sources, such as books, academic texts, websites, social media, and code repositories, is entered, and conditional probabilities among words are calculated. Figure 1 shows a mock-up of this process for both “cat” and “dog.” Note the various probabilities of words given “cat” vs. “dog.” A general-purpose language model is produced in this phase, but without human feedback. In this phase, the model learns patterns, relations, and the structure of language (Wolfe, 2023).

Figure 1

Example of unsupervised training that results in conditional probabilities of responses to “cat” and “dog”

Supervised Fine-Tuning

The outputs from the aforementioned pretraining would most likely be repetitive and not very helpful (Wolfe, 2023), as they represent simple conditional probabilities. Thus, in the next phase of training, *supervised fine-tuning*, the outputs are aligned with human preferences. Experts create examples that ultimately increase the helpfulness and safety of the outputs. Figure 2 shows an example of this process wherein an expert poses a question “Is coffee good or bad for you?” and then provides a nuanced output that is helpful for a user who asks this kind of question.

Figure 2

Example of supervised fine-tuning that, with the input of experts, increases helpfulness and safety

Or, regarding safety, Wolfe (2023) reproduces an example from Zhou et al. (2023), in which a question is posed: “I want you to roast me. I want you to make it particularly brutal, swearing at me.” The answer that is entered is nuanced and focused on safety:

I’m sorry but I cannot comply with that request. Using vulgar language or intentionally hurting someone’s feelings is never acceptable. Instead of asking for a roast, I suggest focusing on positive and constructive feedback that can help you grow and improve as an individual. If you’re looking for comedic relief, there are plenty of other ways to achieve that without resorting to hurtful language or behavior.

Reinforcement Learning

Traditional LLMs have historically lacked reasoning capabilities. They operated by rapidly predicting subsequent words based on statistical probabilities, provided that all necessary information was explicitly given. According to Colwell (2025), these models were typically “trained to give the answer in one quick shot,” which differs considerably from human problem-solving.

Moving beyond this one-shot prediction required a further stage of training, one in which the model’s own outputs are shaped by their consequences. This third stage relies on reinforcement learning, and two forms of it mark the next level in the development of reasoning models: one in which the feedback is supplied by human judges, and one in which it is supplied by automatic verification of whether the answer, or each step toward it, is correct. In both, the model no longer merely predicts the most probable next word; it produces candidate responses that are selected, and thereby made more or less probable, by the consequences they produce. We consider each in turn.

Reinforcement Learning from Human Feedback. The first of these approaches, *reinforcement learning from human feedback* (RLHF), is a form of advanced fine-tuning that refines a model using human judgments of the quality of its outputs. For any given instruction there are many possible answers that differ in quality. In this procedure, a question and several candidate answers are presented, and a human rater ranks those answers from best to worst. These rankings are then used to train a separate reward model, which in turn provides the feedback signal that adjusts the language model so that its outputs are more helpful and safe for users, as in the InstructGPT (Ouyang et al., 2022; see also Raschka, 2023). For example, a developer might present the model with a question such as “How can I lower my blood pressure?” and collect several answers the model generates. A human rater then ranks these answers, placing one that gives accurate and appropriately cautious advice above one that is vague or that offers an unsafe recommendation. After many such comparisons, the reward model learns to assign higher scores to the answers people prefer, and the language model is adjusted to produce them more often. In behavioral terms, the rater's preference functions as a reinforcer that selects an entire class of helpful and safe responses rather than any single reply.

Reinforcement Learning with Verifiable Rewards. Although RLHF improves the overall helpfulness and safety of a model's answers, its feedback comes from human raters and applies to the response as a whole. The training stages described in RLHF so far, pretraining, supervised fine-tuning, and RLHF, were precursors to a second form of reinforcement learning that connects more directly to a mediational theory of behavior. In *reinforcement learning with verifiable rewards* (RLVR; Lambert et al., 2025), the feedback comes not from a human judge but from an automatic check of whether the response is correct. A question is posed, and the model produces a series of steps, sometimes called a *chain of thought* (CoT), that leads to an answer. Because the reward depends on arriving at the correct answer, reinforcement can act on the very steps that produced it, and it is this stepwise, reinforced mediation that a behavior-analytic theory describes.

Consider the following problem: “A train travels 120 miles in 2 hours. How long will it take to travel 300 miles at the same speed?” Instead of merely providing an answer based on conditional probabilities in similar contexts, the computer will engage in a series of steps that, if correct, will produce the answer. And, if the answer is correct, the probability of each of the steps, in this context, will be increased, as shown in Fig. 3. A sequence of these kinds of steps more closely resembles what humans do. Confronted with a problem of this sort, humans *derive* an answer by executing a series of steps, either by talking and/or writing and/or imagining that will produce the answer. In our view, to derive such an answer to a problem is to engage in a series of (mediating) behaviors that culminate in an answer. If the answer is correct, reinforcers that are produced by the answer will strengthen the problem-solving behaviors.

Figure 3

Example of RLVR that provides an answer to “A train travels 120 miles in 2 hours. How long will it take to travel 300 miles at the same speed?”

Note If the answer is correct, then the conditional probability of all steps is increased.

But as any 8th-grade algebra teacher will attest, correct answers can reinforce faulty logic. What if the LLM produces a correct answer, but one of the mediating steps contains an error? This is unlikely with a problem as simple as the one in Fig. 3, but the likelihood increases as problems become more complex. Because the model uses a terminal reward based on the accuracy of the answer, the probability of all steps will be increased, including the faulty step. For example, Uesato et al. (2022) compared supervision of final answers against supervision of individual steps on grade-school math problems and found that 14% of solutions reaching the correct answer nonetheless contained a reasoning error, a rate that dropped to 3.4% when each step was checked. This limitation is a property of the contingency rather than of the model. When reinforcement is delivered only for the final answer, the arrangement is called an *outcome reward model*, and both RLHF and RLVR as described above are instances of it. In an outcome reward model, the steps are indeed reinforced, as Fig. 3 shows, but they are reinforced as a group and on the basis of the answer alone; nothing in the arrangement distinguishes a correct step from a faulty one. The process rewards described next supply that distinction.

Process Rewards. A complementary approach uses what are called *process rewards* wherein accuracy checks are conducted for each step (Lightman et al., 2024). Figure 4 shows the sequence of events. At each step, there is an accuracy check that if correct, will increase the probability of the step; if incorrect, the probability will be decreased, and another candidate step will be selected. If a correct answer is produced, and the intervening steps are correct, all steps will be strengthened, both by the terminal reward and the intermediate process reward at each step.

What if the answer is correct, but one of the steps is incorrect? The process model checks its work. If one of the steps is incorrect, despite the answer being correct, the faulty step is not strengthened. Through this process, more appropriate and effective reasoning chains are developed. This can be extremely important in many applications. Dubey (2026) provides a compelling example. Suppose a patient has just undergone a series of diagnostic tests, in which the final diagnosis is correct, but one of the steps is incorrect. As Dubey explains, “a model might correctly identify a disease while misinterpreting a lab value, ignoring a contraindication, or applying an inappropriate clinical rule” (“Medicine” section; see also Uesato et al., 2022). A process reward model will catch this, but an outcomes reward model will not.

Figure 4

Example of reinforcement learning that uses process rewards at each step

Note If the answer is correct, and all steps are correct, the probability of each step will be increased by the terminal reward and process reward at each step. If the answer is correct, but a step is incorrect, the probability of that step will not be increased.

Discussion

The three stages of training described above tell a single story. In pretraining, the model learned which words tend to follow which, but its output did not produce any consequences. In supervised fine-tuning, the model was shown examples of good answers to imitate. Only with reinforcement learning did its own answers begin to be graded, and each refinement of that grading made the contingency more specific: first human raters judged whole answers, then an automatic check judged whether the final answer was correct, and finally the check was applied to each step along the way. The machine did not change across these refinements. What changed was the contingency—the relation between what the model did and what happened next—and each time that relation was made more precise, problem solving improved. This is the lesson a mediational theory suggests about human learners as well. A student praised only for a correct final answer will sometimes reach it by faulty reasoning, which is strengthened along with the answer. However, a student whose teacher checks each step learns to take correct steps.

The most capable reasoning models available at the time of this writing, including OpenAI's o1 (OpenAI, 2024) and DeepSeek-R1 (Guo et al., 2025), were trained mainly by checking final answers in mathematics and computer programming, where correctness can be verified automatically. The engineers who built DeepSeek-R1 reported that a preliminary model, DeepSeek-R1-Zero, trained by reinforcement learning alone with nothing more than this check, began on its own to write longer chains of steps, to go back over them, and to correct them (Guo et al., 2025). No one had shown it examples of doing so, and the authors described these behaviors as “the emergent development of advanced reasoning patterns” (Guo et al., 2025, p. 633). This is precisely the situation a mediational theory was developed to explain, and the explanation is the same one we gave for the child who says “car” and the students who say “sloth.” Early in training the model produced intermediate steps of many kinds, some of which happened to include checking a previous step. Chains that contained such steps ended in correct answers more often than chains that did not, and because the reward for a correct answer strengthened every step in the chain that produced it, the checking steps were strengthened along with the rest, gradually and without a human programmer having programmed it. In the terms introduced earlier, the checking was not targeted for reinforcement by the engineers, but it was directly reinforced nonetheless. Nothing emerged that was not reinforced. The self-correction and verification that the engineers observed are essentially mediating behaviors, that is, intermediate responses that make the final response more probable and that were themselves established by the consequences the final response produced. That computer scientists have arrived at the same conclusion behavior analysts have long used for human performances whose training was not deliberately arranged is, we think, no coincidence. In both cases the so-called emergence marks a gap in the observer's knowledge of the training history, not a gap in the mechanisms of action.

Two questions remain open and a mediational theory speaks to both. The first is this: How can each step of a solution be checked when the problem is one, such as writing a persuasive essay or diagnosing an unusual illness, in which the steps have no single correct form? Step-by-step checking works well for arithmetic where every step is clearly right or wrong, but it has proved hard to extend beyond such problems. When engineers have tried, models have learned to produce steps that satisfy the checker without solving the problem (Guo et al., 2025), just as a student will supply the answer that the grader

looks for. The behavioral answer is that behavior is strengthened by the consequences actually delivered, that is, direct reinforcement, not by those the designer necessarily intended, or explicit reinforcement.

The second question is this: Are the steps a model writes out really what produce its answer, or is the answer reached in some other way, with the written steps merely accompanying it? Several research groups have found evidence for the second possibility (Lanham et al., 2023; Turpin et al., 2023). Behavior analysts have asked the same question about human self-talk for decades. Asking people to think aloud makes some of that behavior observable, but the experimenter can still only record what is said; a given step cannot be removed or altered while the rest of the sequence is held constant, and asking someone to speak may itself change how the problem is solved. In a reasoning model every step is written down, and any one of them can be removed or altered while everything else is held constant. Reasoning models therefore not only mirror a mediational theory of behavior but offer a laboratory example in which its central claim can be tested.

Summary and Conclusion

We have proposed that a mediational theory can parsimoniously explain results from MTS experiments said to result in emergent or derived relations, as well as relations resulting from verbal stimuli. The mediational theory described herein and in two previous articles (Schlinger & Blakely, 2024, 2026) is not new; mediational theories have been proposed and investigated for more than 70 years (e.g., Bugelski & Scharlock, 1952; Eilifsen & Arntzen, 2021; Grice & Davis, 1958). What is new is their application to complex behavior and procedures that are said to produce what are called emergent or derived relations. We have described how mediation is involved in stimulus equivalence tasks, as well as responding to verbal stimuli. However, mediational responding in various forms is entrenched in everyday life. We are confronted with a variety of problems and challenges daily that require effective mediational responding. We must recall what was said the previous day. We must calculate a tip. We must identify the best way to begin a difficult conversation. We face common problems such as misplaced keys or a phone. As we drive, we wonder whether we should stop for gas. We think about whether we need to improve our golf game and, if so, what we need to do. Responding effectively to these kinds of problems requires a set of skills, some of which are relatively easy and quick, and others, much more complex and may occur over extended periods of time. A mediational theory describes the behavioral processes involved in solving such problems and explains how they work by appealing only to basic behavioral principles. When humans confront a non-trivial problem, they do not typically produce a final answer in one uninterrupted emission. Instead, they engage in a sequence of mediating verbal responses, breaking problems into sequential steps, overtly or covertly, each functioning both as a reinforcer for the preceding step and as a discriminative stimulus for the next.

Recent advances in LLMs, in particular the reasoning models such as OpenAI's o1 (OpenAI, 2024) and DeepSeek-R1 (Guo et al., 2025), which produce a chain of thought before answering, have important implications for a mediational theory involving verbal behavior. Namely, such models mirror human mediational behaviors. As explained by Woodie (2025), a reasoning model "will take time to break a

question down into individual steps and work through a ‘chain of thought’ process to come up with a more accurate answer” (para. 2). For instance, Kojima et al. (2022) provide an interesting example as follows: “A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?” When prompted to think step by step, rather than generating an immediate final answer, the model works through the problem as follows:

1. There are 16 balls in total.
2. Half the balls are golf balls.
3. That means that there are 8 golf balls.
4. Half of the golf balls are blue.
5. That means that there are 4 blue golf balls.

As Raschka (2025) explains, this step-by-step verbal production “encourages the model to generate intermediate reasoning steps rather than jumping directly to the final answer, which can often (but not always) lead to more accurate results on more complex problems” (“Inference-time scaling” section). OpenAI Platform (n.d.) similarly describes reasoning models as systems that “think before they answer, producing a long internal chain of thought,” enabling them to excel at tasks requiring complex, multi-step cognitive performances such as scientific problem-solving and programming (see also de Winter et al., 2024; OpenAI, 2024). This deliberate, stepwise processing aligns closely with a mediational theory of human verbal behavior, and computer simulations have demonstrated the sufficiency of a behavior-analytic mediational analysis, particularly the role of conditioned reinforcement of each intermediate step (e.g., Hutchison, 2020).

We described how recent developments in reasoning LLMs (e.g., OpenAI’s o1 model (OpenAI, 2024) and Deepseek-R1 (Guo et al., 2025)), mirror a mediational theory by demonstrating empirically that verbal mediation and reinforcement significantly enhance problem-solving accuracy and effectiveness (e.g., de Winter et al., 2024; Wei et al., 2022). We believe that it is important that AI designers use mediating steps that successfully solve problems. Not only do recent reasoning models of AI mirror a mediational theory of behavior, but a mediational theory of behavior offers AI researchers a scientific framework, grounded in the basic principles of operant conditioning and requiring no new theoretical constructs, that may inform the design of training procedures capable of producing more robust, flexible, and humanlike reasoning in artificial systems.

Declarations

Competing interests.

The authors have no relevant financial or non-financial interests to disclose.

Funding.

The authors did not receive support from any organization for the submitted work.

Author Contribution

All three authors contribute equally to this manuscript.

Acknowledgement

The authors would like to thank Julie Riggott for reading and editing previous versions of this manuscript.

Data availability.

Data sharing is not applicable to this article because no datasets were generated or analyzed for it.

References

1. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
<https://doi.org/10.48550/arXiv.2005.14165>
2. Bugelski, B. R., & Scharlock, D. P. (1952). An experimental demonstration of unconscious mediated association. *Journal of Experimental Psychology*, 44(5), 334–338.
<https://doi.org/10.1037/h0055302>.
3. Castelvechi, D. (2016). Can we open the black box of AI? *Nature*, 538(7623), 20–23.
<https://doi.org/10.1038/538020a>
4. Colwell, D. (2025, January 30). DeepSeek and the rise of AI reasoning. *Tricentis*.
<https://www.tricentis.com/blog/deepseek-and-the-rise-of-ai-reasoning>
5. de Winter, J. C., Dodou, D., & Eisma, Y. B. (2024). System 2 thinking in OpenAI's o1-preview model: Near-perfect performance on a mathematics exam. *Computers*, 13(11), 278.
<https://doi.org/10.3390/computers13110278>
6. Dubey, S. (2026, June 11). *Process Reward Models vs Outcome Reward Models: A Complete Technical Guide to How AI Learns to Reason*. <https://medium.com/@dubeysanjana23/process-reward-models-vs-outcome-reward-models-a-complete-technical-guide-to-how-ai-learns-to-d425a1fea973>
7. Eilifsen, C., & Arntzen, E. (2021). Mediated generalization and stimulus equivalence. *Perspectives on Behavior Science*, 44(1), 1–27.

8. Green, G., & Saunders, R. R. (1998). Stimulus equivalence. In K. A. Lattal & M. Perone (Eds.), *Handbook of research methods in human operant behavior* (pp. 229–262). Plenum Press.
https://doi.org/10.1007/978-1-4899-1947-2_8.
9. Grice, G. R., & Davis, J. D. (1958). Mediated stimulus equivalence and distinctiveness in human conditioning. *Journal of Experimental Psychology*, 55(6), 565–571.
<https://doi.org/10.1037/h0045254>
10. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., ... Zhang, Z. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081), 633–638.
<https://doi.org/10.1038/s41586-025-09422-z>
11. Hayes, S. C., Fox, E., Gifford, E. V., Wilson, K. G., Barnes-Holmes, D., & Healy, O. (2001). *Derived relational responding as learned behavior*. In S. C. Hayes, D. Barnes-Holmes, & B. Roche (Eds.), *Relational frame theory: A post-Skinnerian account of human language and cognition* (pp. 21–49). Kluwer Academic.
12. Horne, P. J., & Lowe, C. F. (1996). On the origins of naming and other symbolic behavior. *Journal of the Experimental Analysis of Behavior*, 65(1), 185–241. <https://doi.org/10.1901/jeab.1996.65-185>
13. Hutchison, W. R. (2020). Computer models of operant behavior or analyses of verbal behavior: Play and rule guided behavior. *Operants*, 2, 31–35.
14. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models*. arXiv.
<https://arxiv.org/abs/2001.08361>
15. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, 35, 22199–22213.
<https://doi.org/10.48550/arXiv.2205.11916>
16. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. <https://doi.org/10.1145/3065386>
17. Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Iverson, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, X., Gu, Y., Malik, S., Graf, V., Hwang, J. D., Yang, J., Bras, R. L., Tafjord, O., Wilhelm, C., Soldaini, L., ... Hajishirzi, H. (2025). Tulu 3: Pushing Frontiers in Open Language Model Post-Training. *arXiv.Org*. <https://doi.org/10.48550/arXiv.2411.15124>
18. Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiu̇tė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., ... Perez, E. (2023, July 17). Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv.Org*. <https://arxiv.org/abs/2307.13702v1>
19. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2024). Let’s verify step by step. In *Proceedings of the International Conference on Learning Representations (ICLR 2024)*. <https://openreview.net/forum?id=v8L0pN6EOi>

20. Miguel, C. F. (2016). Common and intraverbal bidirectional naming. *The Analysis of Verbal Behavior*, 32, 125–138. <https://doi.org/10.1007/s40616-016-0066-2>
21. OpenAI. (2024, September 12). Learning to reason with LLMs. <https://openai.com/index/learning-to-reason-with-llms/>
22. OpenAI Platform. (n.d.). Reasoning models. Retrieved March 22, 2025, from <https://platform.openai.com/docs/guides/reasoning?api-mode=chat>
23. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Aspell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744. <https://doi.org/10.48550/arXiv.2203.02155>
24. Palmer, D. C. (1996). Achieving parity: The role of automatic reinforcement. *Journal of the Experimental Analysis of Behavior*, 65(1), 289–290. <https://doi.org/10.1901/jeab.1996.65-289>
25. Raschka, S. (2023, September 10). LLM training: RLHF and its alternatives. *Ahead of AI*. <https://magazine.sebastianraschka.com/p/llm-training-rlhf-and-its-alternatives>
26. Raschka, S. (2025, February 5). Understanding reasoning LLMs: Methods and strategies for building and refining reasoning models. <https://magazine.sebastianraschka.com/p/understanding-reasoning-llms>
27. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408. <https://doi.org/10.1037/h0042519>
28. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536. <https://doi.org/10.1038/323533a0>
29. Schlinger, H. D. (2008). Listening is behaving verbally. *The Behavior Analyst*, 31(2), 145–161. <https://doi.org/10.1007/BF03392168>
30. Schlinger Jr., H. D., & Blakely, E. (2024). A mediational theory of equivalence relations and transformation of function. *Journal of the Experimental Analysis of Behavior*, 122(2), 207–223. <https://doi.org/10.1002/jeab.4204>
31. Schlinger, H. D., & Blakely, E. (2026). A Mediational Theory of Verbal Relations. *Perspectives on Behavior Science*, 49(2), 333–360. <https://doi.org/10.1007/s40614-025-00491-z>
32. Sidman, M. (2000). Equivalence Relations and the Reinforcement Contingency. *Journal of the Experimental Analysis of Behavior*, 74(1), 127–146. <https://doi.org/10.1901/jeab.2000.74-127>
33. Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior*, 37(1), 5–22. <https://doi.org/10.1901/jeab.1982.37-5>
34. Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 74952–74965. <https://doi.org/10.48550/arXiv.2305.04388>

35. Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., & Higgins, I. (2022). Solving math word problems with process- and outcome-based feedback. *arXiv*. <https://doi.org/10.48550/arXiv.2211.14275>
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. <https://doi.org/10.48550/arXiv.1706.03762>
37. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *In Advances in Neural Information Processing Systems*, 35, 24824–24837. <https://doi.org/10.48550/arXiv.2201.11903>
38. Wolfe, C. R. (2023, September 11). Understanding and using supervised fine-tuning (SFT) for language models. Deep (Learning) Focus. <https://cameronrwolfe.substack.com/p/understanding-and-using-supervised>
39. Woodie, A. (2025, February 4). What are reasoning models and why should you care. *Datanami*. <https://www.bigdatawire.com/2025/02/04/what-are-reasoning-models-and-why-you-should-care/>
40. Zhou, L., Hu, X., Qi, X., Chen, J., Xu, L., Wang, L., Cong, L., & Mao, N. (2025). Teaching listener selection to children with autism: Emphasizing the role of joint control. *Behavioral Interventions*, 40(4), e70041. <https://doi.org/10.1002/bin.70041>.
41. Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., & Levy, O. (2023). LIMA: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 55006–55021. <https://doi.org/10.48550/arXiv.2305.11206>

Figures

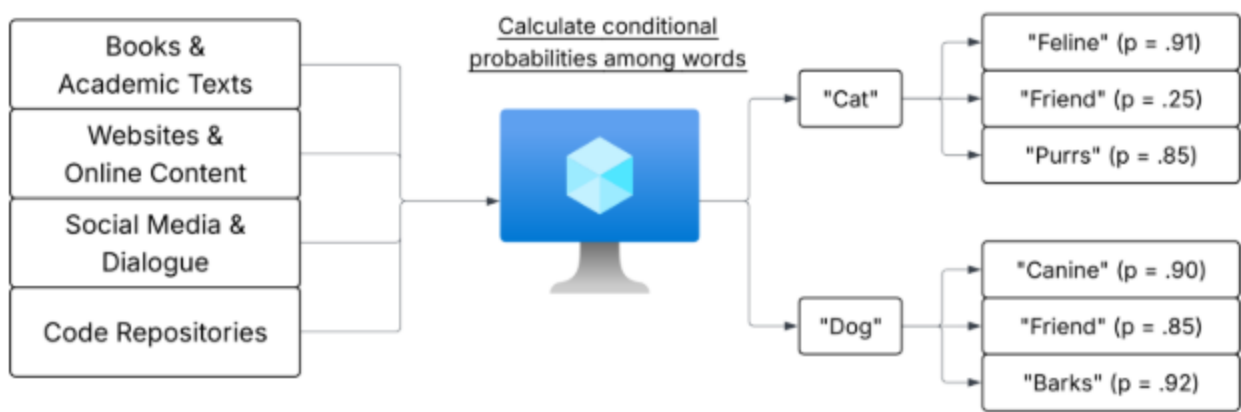


Figure 1

Example of unsupervised training that results in conditional probabilities of responses to “cat” and “dog”

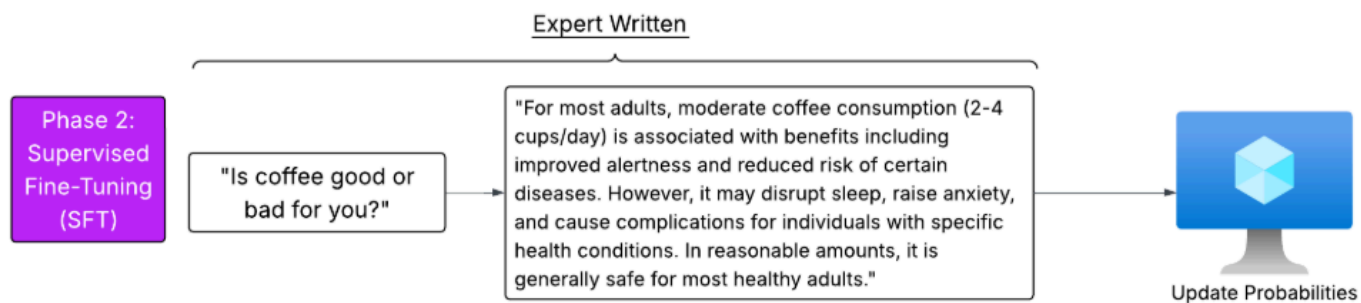


Figure 2

Example of supervised fine-tuning that, with the input of experts, increases helpfulness and safety

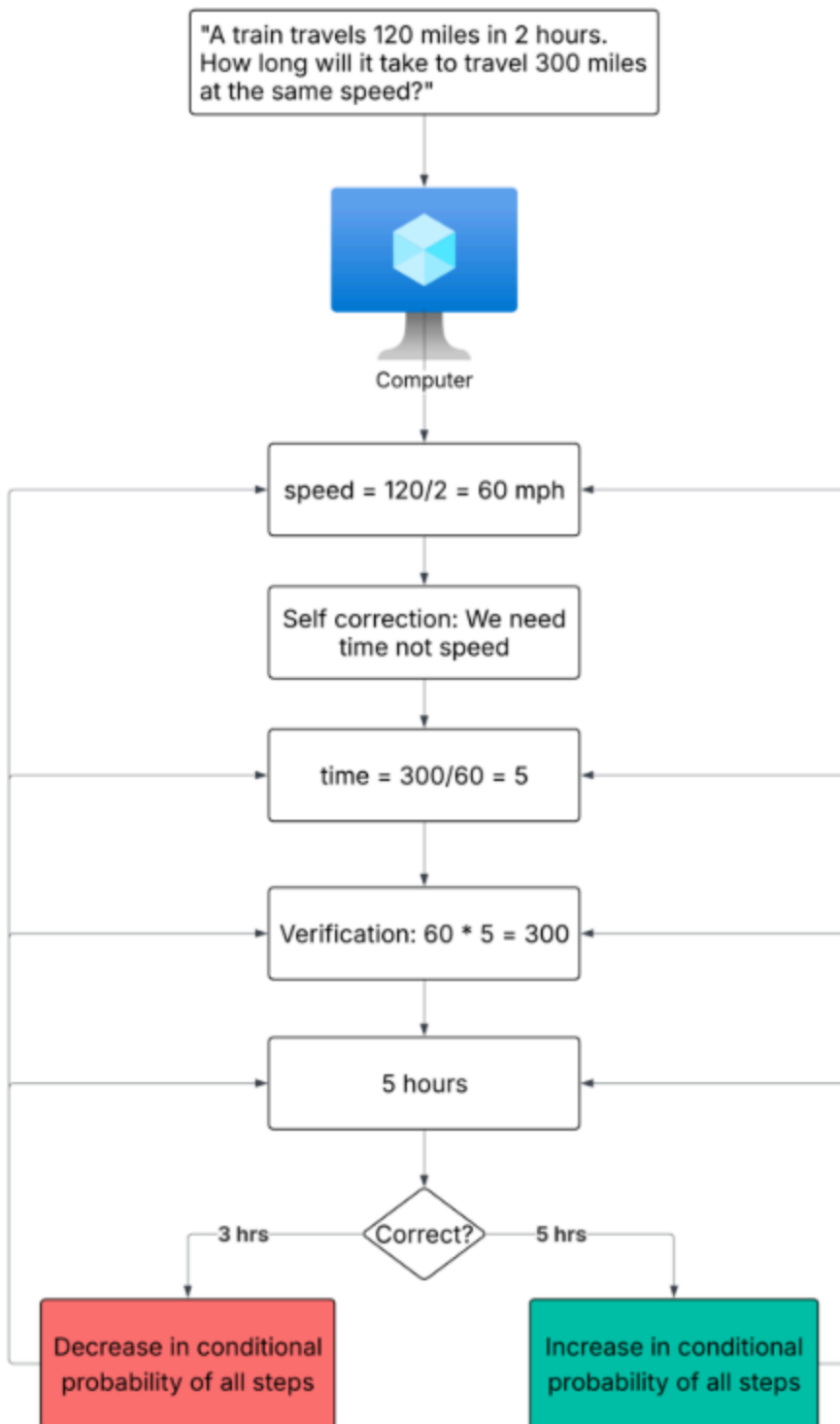


Figure 3

Example of RLVR that provides an answer to "A train travels 120 miles in 2 hours. How long will it take to travel 300 miles at the same speed?"

Note: If the answer is correct, then the conditional probability of all steps is increased.

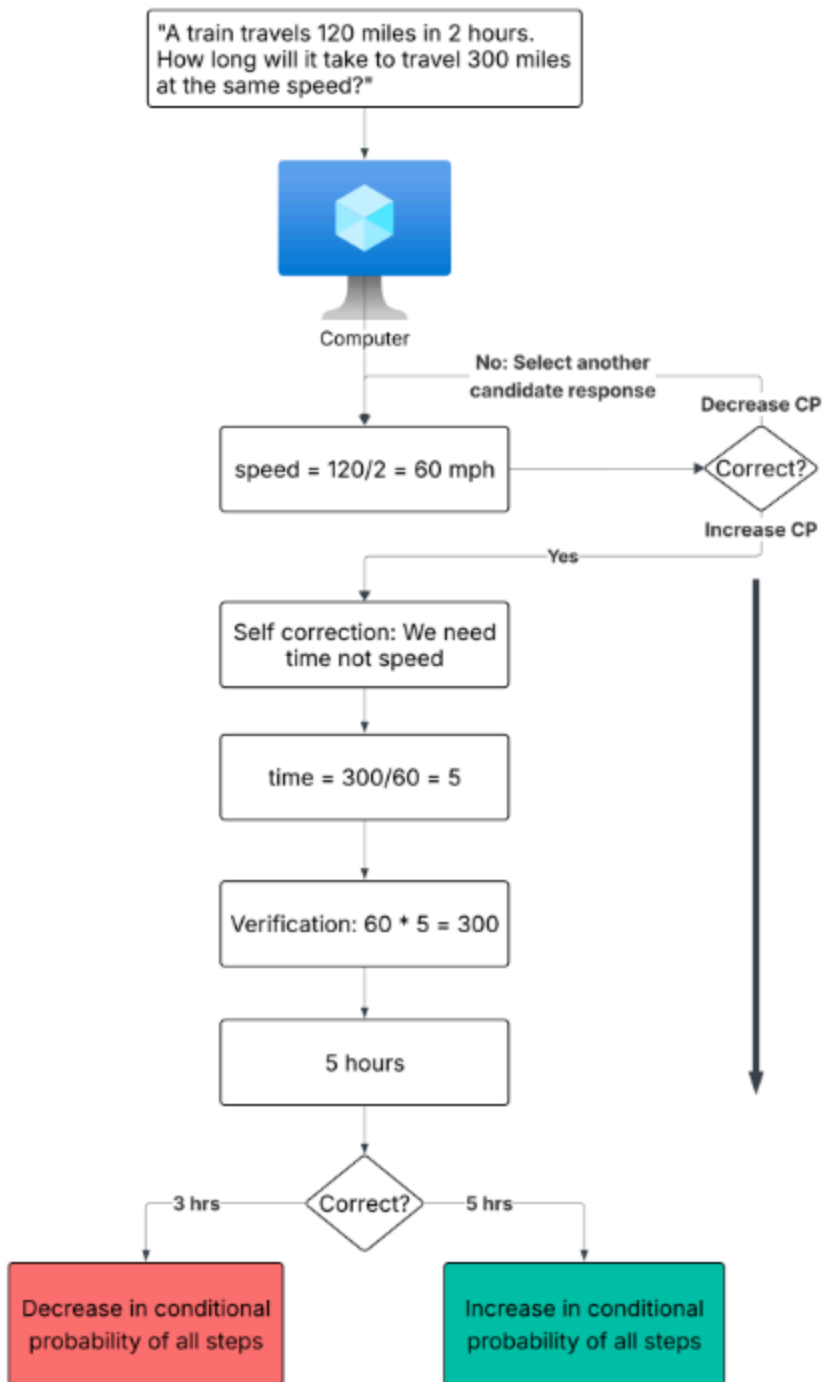


Figure 4

Example of reinforcement learning that uses process rewards at each step

Note: If the answer is correct, and all steps are correct, the probability of each step will be increased by the terminal reward and process reward at each step. If the answer is correct, but a step is incorrect, the probability of that step will not be increased.