# Thinking fast, slow, and everywhere in between in humans and language models

**Ben Prystawski**
Department of Psychology
Stanford University
benpry@stanford.edu

**Noah D. Goodman**
Departments of Psychology and Computer Science
Stanford University
ngoodman@stanford.edu

## Abstract

How do humans adapt how they reason to varying circumstances? Prior research has argued that reasoning comes in two types: a fast, intuitive type and a slow, deliberate type. Are these the only options, or can people adjust their reasoning continuously by trading off speed and accuracy within individual reasoning steps? We investigate this possibility in an experiment where participants were trained on relationships between local variables in a simple causal model, then asked to make predictions about all pairs of variables. Participants in one condition had a 5-second time limit. We found main effects of time pressure and locality, but only a small interaction in the direction opposite to our hypothesis. We present a process-level model of this phenomenon using early readouts from transformer language models. Our findings are consistent with people reasoning step by step, but accepting a higher error rate at each step under time pressure.

**Keywords:** reasoning; dual process theory; speed-accuracy tradeoff; language models

## Introduction

Human reasoning is remarkably flexible. Thinking through a problem step by step can enable people to make much better inferences than they would make directly. People have developed a range of tools for thinking that improve our ability to solve problems, including elaborate thought experiments and processes of deduction (Dennett, 2013; Shepard, 2008). We can even reason under difficult circumstances, like impending deadlines and noisy rooms, though at some cost to accuracy. Somehow, the mind implements a tradeoff between the speed of reasoning and the accuracy of results. How do we do it?

Dual process theory, a popular framework in the study of reasoning, answers this question by positing that humans can make use of two types of reasoning: a fast, intuitive type often called type 1 and a slow, deliberate type called type 2 (Evans, 2003; Kahneman, 2011; Wason & Evans, 1974). These processes may be framed as distinct mental systems, or different ways of using the same system. Though many variants of dual process theory have been proposed, they all characterize the mind as reasoning in two distinct ways (Evans, 2012).

Some of the evidence for dual process theory comes from experiments where time pressure is used to constrain people's ability to think. People make predictable errors on tasks where type 1 and type 2 thinking lead to different answers, and time limits make those errors more common. In syllogistic reasoning experiments, giving participants a time limit makes them more likely to rely on how plausible they find the *conclusion* of a syllogism, rather than its logical validity (Evans & Curtis-Holmes, 2005). Similar patterns have been found in reasoning about conditional statements (Evans et al., 2009) and the Wason card selection task (Roberts & Newton, 2001; Wason, 1968). A key finding from these studies is that careful step by step reasoning takes time.

A similar bias toward the believability of the conclusions has been found in large language models when prompted with reasoning tasks (Lampinen et al., 2024). This finding builds upon a line of work on "chain of thought" reasoning in language models. Prompting language models to think step by step reliably improves their performance on tasks that humans need to think deliberately to do well on, like math and logic (Nye et al., 2021; Wei et al., 2022), and decreases performance on tasks where deliberate thinking makes humans worse (Liu et al., 2024). Chain of thought reasoning has therefore been considered as a computational analogue to human reasoning (e.g., Prystawski et al., 2023).

When participants exhibit belief bias, they substitute a difficult process of thinking for an easier intuitive answer. Not all tasks have obvious intuitive answers, though. How, then, might people respond to time pressure in reasoning tasks where there is not an obvious intuitive answer? One possibility is that under time pressure, people lack the time to reason deliberately, so they either guess randomly or rely on spurious patterns.

Another possibility is that people attempt to reason step by step, but spend little time on each step. Many facets of cognition, including perception, memory, and decision-making, exhibit continuous relationships between speed and accuracy (Wickelgren, 1977). People might then reason quickly by working through the same series of steps they otherwise would, but accepting a higher rate of error at each step.

In this paper, we study how people trade off between speed and accuracy in a reasoning task inspired by a task used to test the effectiveness of reasoning step by step in language models. We train participants on local dependencies in a causal model with a chain structure, then ask them to make predictions about all pairs of variables in the model. We manipulate people's ability to reason through time pressure: assigning participants to either a speeded condition with a 5-second time limit or an unspeeded condition with no time limit.

We initially hypothesized that participants in the speeded condition would show a particularly large drop in accuracy



In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* 

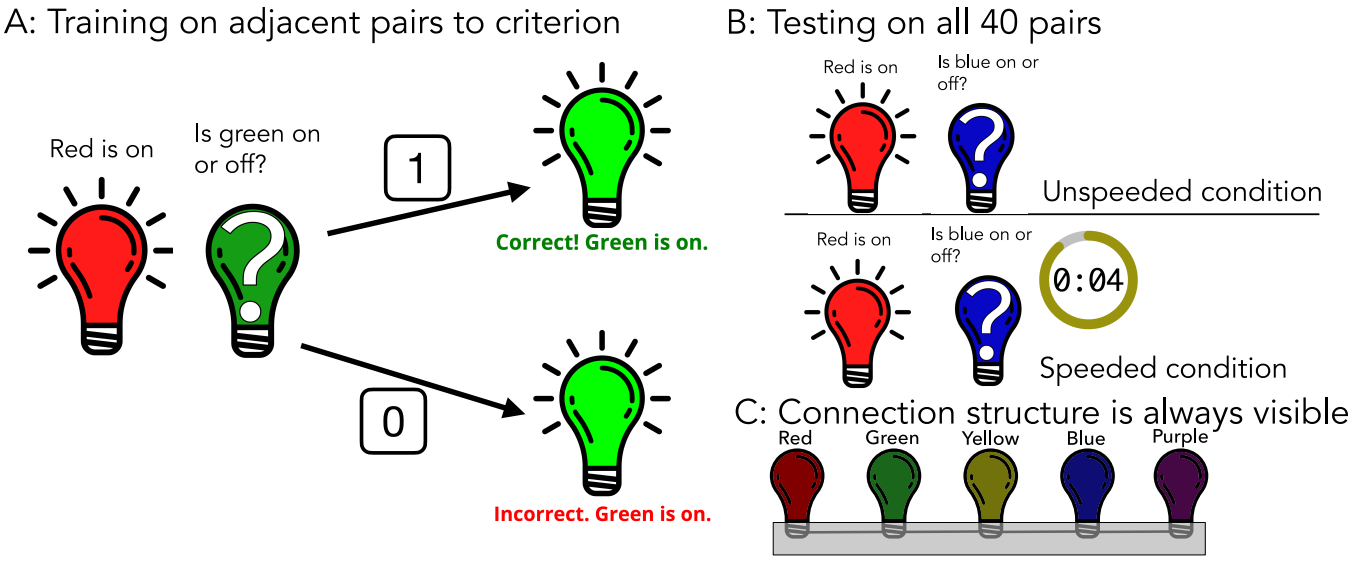


Figure 1: Overview of our experimental setup. A: Participants are first trained to criterion to predict whether a light is on or off given the value of an adjacent light. B: Next, they are asked to make predictions about 40 possible combinations of observed light, observed value, and target light. Depending on the condition, participants may have a 5-second time limit when making predictions. C: Participants can see how the lights are connected throughout the experiment.

for non-local inferences. This would be consistent with a dichotomous account of reasoning, where participants cannot reason step by step at all under time pressure. Instead, we found that performance dropped similarly for inferences of all distances in the speeded condition. These findings are consistent with participants trading off between speed and accuracy at each step, rather than switching between two distinct modes of thinking. To make this account more concrete, we train transformer language models on the same task and show that reading out from intermediate layers can provide a process-level account of this pattern of results.

## Methods

We showed participants samples from a causal model with Boolean-valued variables, framed as lights on a machine being on or off. Participants first completed a local training phase in which they learned the dependencies between adjacent lights. After learning these local dependencies, they completed a test phase where they made predictions about all pairs of lights, including those they did not see together in training. Participants were assigned to either the speeded condition, where they had to make test phase predictions within 5 seconds, or the unspeeded condition, where there was no time limit. Figure 1 shows an overview of our experimental setup. The preregistration can be found at https://osf.io/fy2a4. Code and data can be found at https://github.com/benpry/thinking-in-chains

### Participants

We recruited 184 participants through the crowdsourcing platform Prolific. Participants received $2 for their participation, with a potential bonus of $1 based on their performance. They took a median of 14 minutes to complete the experiment.

We lost data from 7 participants due to technical problems. Participants were excluded if they indicated in the post-experiment survey that they were colorblind (2) or that they took notes (6). We also excluded participants who had a response time of less than 100 milliseconds for at least 10 of the 40 test phase trials (1) or took 20 repetitions (320 trials) or more to pass the training phase (7). In total, we excluded 15 participants. This left us with 81 participants in each condition.

### Stimuli

Participants viewed the connection structure of the lights at the beginning of the experiment. This structure remained visible throughout both the training and test phases.

The lights and their values (on and off) represent variables in a causal model. The model has five variables arranged in a simple chain structure. Every connection was deterministic. Pairs of lights are either perfectly correlated (i.e. one is always on when the other is on and off when the other is off) or perfectly anti-correlated (i.e. one is always off when the other is on and vice-versa). Deterministic connections were chosen to make the task easy to learn and to ensure that participants' accuracy did not depend on any intrinsic uncertainty in the task. We randomly assigned participants to one of four chains with different arrangements of correlations and anti-correlations between adjacent lights.

## Procedure

When participants began the experiment, they read instructions telling them that they will see a machine with five lights of different colors. They were told that the lights were connected in the manner shown in Figure 1 and that each pair of adjacent lights either supported each other (i.e. were on and off at the same time) or inhibited each other (i.e. one was off when the other was on). The instructions then described the training and test phases. Participants were told that a random trial had been chosen to be the bonus trial and they would earn a bonus of $1 if they got the answer right on the bonus trial. They were also instructed not to take notes.

Participants first completed a learning phase in which they saw pairs of adjacent lights presented one by one. Each pair consisted of one *observed* light that was shown to be on or off and another *target* light whose value participants needed to predict. They received feedback after each prediction telling them whether their prediction was right and what the true value was. This phase repeated in blocks of 16 trials where the participant saw each combination of observed light, observed value, and target light until they predicted at least 15 out of 16 trials correctly in two successive blocks.

Next, participants completed the test phase. In the test phase, they were asked about all possible combinations of observed light, observed value, and target light. There were a total of 40 test trials. Participants in the speeded condition saw a 5-second timer counting down. If they did not submit a response in 5 seconds, the experiment automatically moved to the next trial and their prediction was considered incorrect. In the unspeeded condition, participants had no time limit.

## Noisy step-wise reasoning

This task does not permit a quick, intuitive strategy that leads to performance better than chance. Each light is equally likely to be on or off *a priori*, and no other features of the light predict whether it is on or off in any given trial. If participants switch between deliberate reasoning and no reasoning at all, we might expect performance to resemble the left panel of Figure 2: a small decrease in accuracy with distance when participants have time to think carefully and a dramatic drop to chance for non-local inferences under time pressure. This was our initial hypothesis: time pressure would create a particularly sharp decrease in accuracy for non-local inferences. We thought that the local inferences might show a small decrease in accuracy with time pressure, but we expected this effect to be small since participants were already trained on local inferences.

However, participants might be more successful in the speeded condition if they chain together a sequence of quick, possibly inaccurate predictions at each reasoning step. If they do this, we should expect accuracy to remain above chance for non-adjacent pairs, but decrease steadily with the number of steps between the observed and target lights.

To characterize this process of noisy step-wise reasoning more precisely, we can assume that a reasoner makes an error

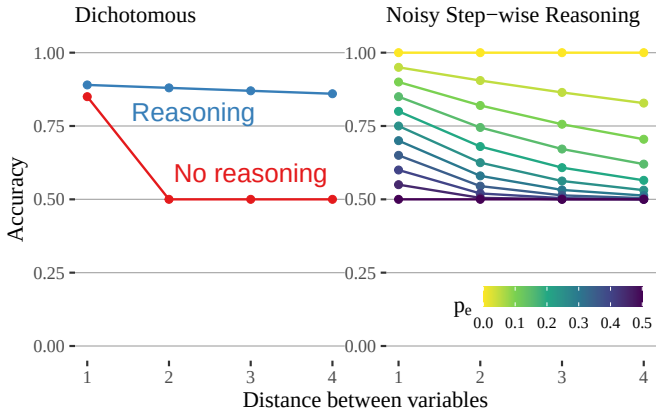


Figure 2: Left: predicted accuracy by distance in the dichotomous account when people can reason step by step (blue) and when they cannot (red). Right: predictions of noisy step-wise reasoning for different step-wise error probabilities.

at any given reasoning step with probability $p_e$, then we can model the number of errors on the overall problem, $N_e$, as a Binomial distribution parameterized by $p_e$ and the number of steps, $n$. Responses in this task are Boolean-valued, so making an even number of errors still leads to the correct answer because the errors cancel each other out. Therefore, the probability of answering correctly is the probability of making an even number of errors:

$$p_{\text{correct}} = p(N_e \text{ is even}) \quad (1)$$

$$N_e \sim \text{Binomial}(n, p_e) \quad (2)$$

The right panel of Figure 2 shows the probability of answering correctly as a function of the distance between variables and the probability of error in an individual reasoning step, interpolating between perfect ($p_e = 0$) and chance ($p_e = 0.5$). We might think of the effect of time pressure as increasing $p_e$, producing the changes to overall accuracy displayed in the figure.

# Results

As a manipulation check, we first calculated participants' response times in each condition. In the last two blocks of the training phase, participants took a median of 2.50s (interquartile range [IQR]: 2.26s) to respond in the unspeeded condition and 3.22s (IQR: 2.67s) to respond in the speeded condition. In the test phase, participants took a median of 3.52 seconds (IQR: 3.60s) to respond in the unspeeded condition and a median of 2.46 seconds (IQR: 1.58s) in the speeded condition. Even though most test phase responses in the unspeeded condition took fewer than 5 seconds, the manipulation still induced a significant difference in response times ($p < 0.0001$ according to a permutation test).

We preregistered two main analyses, the results of which we outline below.

### Effects of locality

We tested two main hypotheses: an *accuracy hypothesis* that participants in the speeded condition would demonstrate a steeper decrease in accuracy for non-local inferences than participants in the unspeeded condition, and a *response time hypothesis* that participants in the unspeeded condition would require more time to respond for non-local inferences compared to local inferences. We tested both of these hypotheses using Bayesian mixed-effects models with random intercepts and slopes on locality for each connection structure and participant. We report coefficient estimates ($\beta$) for linear regressions and odds ratios (OR) for logistic regressions. All estimates are reported with 95% credible intervals. We group all comparisons between non-adjacent variables together as "non-local", since participants did not see the variables together in training. In preregistered secondary analyses presented in the next section, we consider the effect of distance among non-local pairs.

We tested the accuracy hypothesis using a logistic regression model with main effects of condition (speeded vs. unspeeded) and locality (local vs. non-local) and an interaction term. We found strong main effects of condition (OR: 7.84 [4.52, 14.50]) and non-locality (OR: 0.31 [0.14, 0.68]), as well as an interaction that went in the opposite of the hypothesized direction: participants in the speeded condition showed a slightly larger decrease in performance for local inferences compared to non-local inferences (OR: 0.53 [0.31, 0.92]). We can understand why this might have occurred by looking at the relationship between distance and accuracy in Figure 3. Participants in the speeded condition perform substantially worse for distance-1 (local) inferences and chance accuracy is 50%, so there is simply less room for performance to fall in the speeded condition.

We tested the response time hypothesis using a linear regression predicting the participants' response times in the unspeeded condition, in seconds, with a fixed effect of locality. We found a significant increase in response times for non-local inferences compared to local inferences ($\beta = 1.72\ [0.56, 2.81]$). Participants took longer to make inferences between variables they had not seen together in training.

### Effect of distance among non-local pairs

While our main hypotheses concern the difference between local and non-local pairs of lights, we also hypothesized that there might be differences among non-local pairs depending on the distance between the observed and target lights. If people reason about non-local pairs of lights by chaining together learned dependencies between local pairs, we would expect them to take longer and make more errors the more steps they need to take. Since there are fewer pairs of lights at greater distances, we group together distances 3 and 4.

First, we analyzed participants' accuracy on pairs of lights with distance 2 and distance 3 or 4 using a Bayesian mixed-effects logistic regression with fixed effects of condition (speeded or unspeeded) and distance (2 or $> 2$) and an interaction between the two. We found no significant interaction between distance and condition (OR: 0.92 [0.57, 1.54]). There was a strong main effect of condition (OR: 4.39 [2.82, 7.13]), as well as a main effect of distance (OR: 0.63 [0.40, 0.999]). Response times were numerically higher for inferences of distance 3 and 4 compared to distance 2, but this difference was not significant ($\beta = 0.67\ [-0.62, 1.87]$).

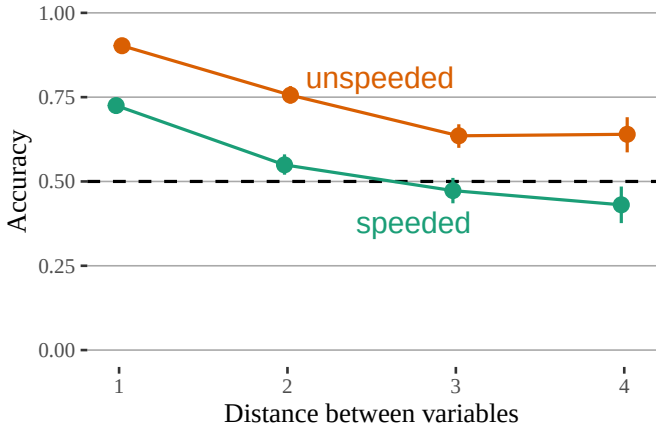


Figure 3: Accuracy by distance between the observed and target variable and condition (speeded vs. unspeeded). Error bars show bootstrapped 95% confidence intervals. Dashed line denotes chance performance.

### Discussion

The results of this experiment, particularly with respect to accuracy, contradicted our initial hypotheses. We expected participants to perform similarly across conditions on the local inferences, and for participants in the speeded condition to perform much worse on non-local inferences. Instead, we found strong main effects of condition and locality and a small interaction in the opposite of the hypothesized direction. Participants did substantially worse in the speeded condition even for the local inferences that they were trained on. Participants in the unspeeded condition performed worse on the more distant inferences as well.

In the unspeeded condition, participants took significantly longer to make non-local inferences compared to local inferences, indicating some additional thinking. Furthermore, the decline in accuracy with distance shown in Figure 3 suggests that participants made more errors the more steps they needed to think through.

The pattern of participant performance in this experiment resembles an increase in the error probability in each step, as visualized in the right panel of Figure 2, more than a dichotomous shift from reasoning to not reasoning. However, we still lack a process-level model of this style of reasoning. What underlying algorithm might produce these trends? In the following section, we draw from research on reasoning in language models to develop such a model.

## Computational analyses

We explore how the noisy step-wise reasoning that explains human performance might be realized algorithmically by training autoregressive transformer language models from scratch on an analogue to our task until they reach a high level of performance. We then mimic the effect of time pressure by reading out predictions from earlier layers of the model at each reasoning step. The models still work through all of the steps, but early readouts mean that less computation is required for each step. This makes it analogous to humans having less time to think through each step.

### Model training and inference

We trained a small version of the GPT-2 architecture (Radford et al., 2019) with 256-dimensional embeddings, 8 layers, and 2 attention heads on a dataset consisting of adjacent pairs of variables and their values. We used the Adam optimizer (Kingma & Ba, 2015) with a learning rate of $5 \times 10^{-5}$ and a linear scheduler. In framing the task as a language modeling problem, we assign a letter name to each light and represent it being on and off with the numbers 1 and 0, respectively. For example, a pair where the red light is on and the green light is off would be written as "A=1\nB=0". The model is trained on a corpus of pairs of adjacent variables and their values formatted in this way, with pound signs separating the samples. This task format closely follows the training setup used by Prystawski et al. (2023).

To match participants' training, we evaluated the model's performance on all the adjacent variable pairs during training. After every gradient step, we produced predictions for all adjacent variable-value pairs and computed the probability the model assigns to the correct answer. We then averaged the probabilities assigned to the correct answers over all adjacent pairs. If the average probability exceeded 90%, we considered the model to have reached criterion and stopped training. Figure 4 shows an overview of our training and inference setup.

We trained five transformers for each causal model in the experiment, each of which used a different random seed. This gave us a total of 800 model predictions.

### Early readouts as an algorithmic model of compute-accuracy trade-offs

Given a trained model, we can elicit predictions for the next token in the standard way: simply applying the language modeling head to the last hidden state corresponding to the last token. Alternatively, we could apply the language modeling head to earlier layers. Reading out from earlier layers requires less compute to make predictions, making it analogous to humans having less time to think.

When predicting the values of variables, we compute the softmax of the logits corresponding to the tokens for 0 and 1. This ensures that the predictions we elicit are true probability distributions over the possible values of the lights, which is particularly important when we read out from earlier layers. Participants in our experiment had perfect knowledge

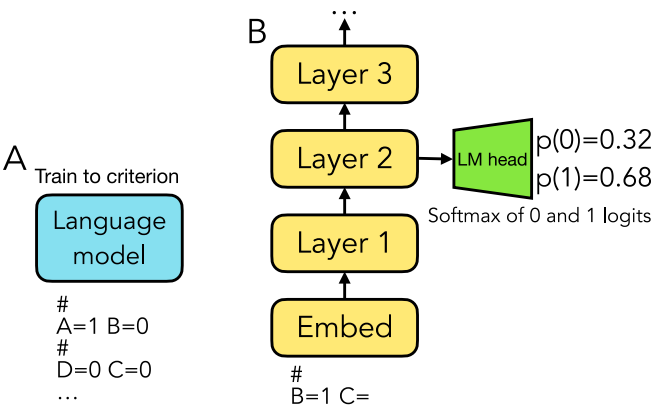


Figure 4: Overview of our model training and inference. A: We train language models to criterion on pairs of adjacent variables. B: We predict conditional probabilities by applying the language modeling head to intermediate states and computing the probabilities of the tokens for 0 and 1.

of how the lights were connected throughout the experiment, and thus always knew the right set of steps to reason through. We recreate this using an estimator that predicts a value for each variable on the path between the observed and target variable one by one, keeping only the most-recently generated value in context. This estimator ensures that the model generates the right intermediate variables to reason through and respects their conditional independence structure. We approximately marginalize over the intermediate variables by sampling 10 sets of intermediate variables from the language model and averaging the conditional probability estimates over those samples.

The probabilities assigned to the correct answers for different intermediate readout layers and distances between observed and target variables are shown in Figure 5. Reading out from earlier layers generally increases the error probability at each reasoning step, and accuracy drops off steadily as a function of distance. There was considerable variability based on the connection structure of the chains. In chains with two anti-correlated pairs of lights, reading out from layer 6 sometimes does just as well or better than reading out from layer 8. Still, the aggregate pattern of results is similar to the trend shown by participants in our experiment. Comparing readouts from layers 4 and 6 produces similar trends to the humans in the speeded and unspeeded conditions, respectively. Like in humans, a Bayesian mixed-effects model fit to our language models' accuracy shows significant main effects of non-locality ($\beta = -0.15$ $[-0.22, -0.08]$) and readout layer ($\beta = 0.15$ $[0.13, 0.17]$), but only a small interaction between them with a credible interval that contains 0 ($\beta = 0.02$ $[-0.001, 0.05]$). Restricting the amount of computation the model can perform at each reasoning step can therefore capture the aggregate pattern of human performance.

## General Discussion

A common view on reasoning posits that people employ two distinct types of reasoning: a fast, heuristic type 1 and a slow, deliberate type 2 (Evans, 2003; Wason & Evans, 1974). This view fits the evidence from tasks where an easy heuristic is available, but has less to say about what to expect when no such heuristic exists. Here, we investigated the relationship between the number of reasoning steps and time pressure in a task where participants predicted the value of one light based on another. Contrary to our original dichotomous hypothesis, we found strong main effects of both distance and condition (speeded or unspeeded), but only a small interaction in the opposite of the hypothesized direction. These results are consistent with a speed-accuracy trade-off in each reasoning step, as we have shown by comparing the pattern of accuracy to an analytic model and trained transformers. Rather than using entirely different strategies in the speeded and unspeeded conditions, participants appeared to rely on a stepwise estimator in both conditions, but tolerated more errors at each reasoning step in the speeded condition.

This work takes an important step toward understanding the range of reasoning strategies people use in tasks without an obvious heuristic available. Many standard tests used to differentiate between type 1 and type 2 reasoning, such as the Cognitive Reflection Test (Frederick, 2005), require participants to reject an intuitively appealing answer and think carefully about the question. In tasks like ours, it is likely clear to participants that reasoning is the only way they can solve the task, so relying on a quick heuristic is not possible.

It may still be surprising that participants showed such a drop in accuracy for local inferences in the speeded condition. If participants can perform three or four noisy reasoning steps within the five-second time limit, why can they not perform a single less-noisy step? While we do not have a definitive answer, it could be that deciding how much thought to put into each step at the beginning of every trial creates too much cognitive overhead, so the best option is to accept a fixed amount of error in each step. This behavior could reflect rational metareasoning (Lieder & Griffiths, 2017).

The existence of compute-accuracy tradeoffs in transformers raises the possibility for inference algorithms that choose what layer to read out predictions from dynamically in order to balance performance on reasoning tasks with the computational cost of inference. Just as humans decide how hard to think in each reasoning step, we could design algorithms that choose how much compute to allocate to each step when presented with a problem. This approach would align with recent work applying the concept of rational metareasoning to language models (De Sabbata et al., 2024).

Although methods exist for eliciting sensible predictions from early hidden states of transformers (Belrose et al., 2023), we did not need to apply them to the models we trained on our task. This is non-obvious. Nothing in the training setup or architecture guarantees that reading out predictions from early layers using the language modeling head should work, yet we find that average accuracy degrades smoothly as we read out from earlier layers. It is possible that residual connections in the transformer architecture lead models to learn to build up predictions gradually over layers, though further research in machine learning interpretability would be valuable to shed light on when and why this happens.

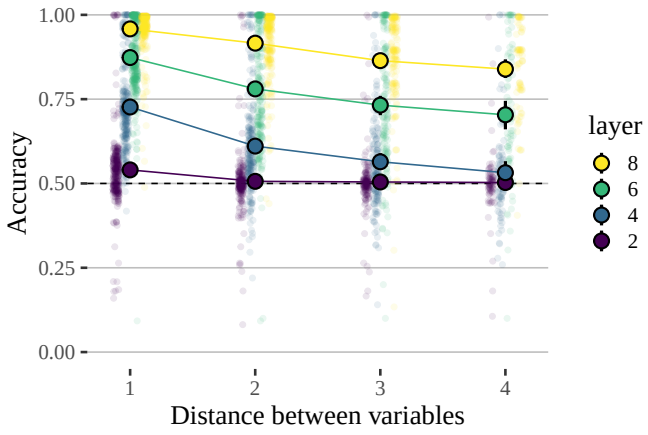


Figure 5: Accuracy by readout layer in transformer language models trained on data that mimics our task. Data is aggregated across 4 chains and 5 training runs per chain with different random seeds. Small points represent individual pairs for a single chain and seed, and large dots show averages.

The task used in this study represents possibly the simplest case of human reasoning: making inferences about a deterministic chain of five variables with a known structure. This approach has several advantages. First, the deterministic nature of the chain makes it relatively easy for humans to learn over the course of an experiment. People often struggle to learn and explicitly reason about probabilities, so a stochastic setting would be less well-suited to an experiment like this (Gigerenzer & Gaissmaier, 2011; Tversky & Kahneman, 1974). Second, the fact that participants are told the connection structure means that we can be reasonably confident that we know which intermediate steps they reason through, so the concept of distance is coherent.

However, the simple nature of our task may limit its ability to distinguish between different models of reasoning. Other models of reasoning, like careful reasoning with early stopping when a time limit is reached, might also account for human behavior in this task.

Future work should consider these additional models, which could require more complex variants of the task that can better distinguish between competing theories. It may be valuable to design a variant of the task where a simple heuristic *is* available, for instance by manipulating the marginal probability of lights being on, to compare more possible accounts of reasoning. Future work should also explore how broadly early readouts from language models can be used as models of reasoning under time pressure.

## Acknowledgments

The authors thank the Stanford Computation and Cognition lab for advice and feedback on this project as it developed. Particular thanks are extended to Daniel Wurgaft for insights on experiment design, as well as to Veronica Boyce, Alexa Tartaglini, and Linas Nasvytis for their valuable comments on a draft of this manuscript.

## References

Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., & Steinhardt, J. (2023). Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

De Sabbata, C. N., Sumers, T. R., & Griffiths, T. L. (2024). Rational metareasoning for large language models. *arXiv preprint arXiv:2410.05563*.

Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. WW Norton & Company.

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in cognitive sciences*, *7*(10), 454–459.

Evans, J. S. B. T. (2012). Dual process theories of deductive reasoning: Facts and fallacies. *The Oxford handbook of thinking and reasoning*, 115–133.

Evans, J. S. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, *11*(4), 382–389.

Evans, J. S. B. T., Handley, S. J., & Bacon, A. M. (2009). Reasoning Under Time Pressure: A Study of Causal Conditional Inference. *Experimental Psychology*, *56*(2), 77–83.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, *19*(4), 25–42.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, *62*(1), 451–482.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus; Giroux.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.

Lampinen, A. K., Dasgupta, I., Chan, S. C., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2024). Language models, like humans, show content effects on reasoning tasks. *PNAS nexus*, *3*(7), pgae233.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological review*, *124*(6), 762.

Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., & Griffiths, T. L. (2024). Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*.

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al. (2021). Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Prystawski, B., Li, M., & Goodman, N. (2023). Why think step by step? reasoning emerges from the locality of experience. *Advances in Neural Information Processing Systems*, *36*, 70926–70947.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Roberts, M. J., & Newton, E. J. (2001). Inspection times, the change task, and the rapid-response selection task. *The Quarterly Journal of Experimental Psychology Section A*, *54*(4), 1031–1048.

Shepard, R. N. (2008). The step to rationality: The efficacy of thought experiments in science, ethics, and free will. *Cognitive Science*, *32*(1), 3–35.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, *185*(4157), 1124–1131.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly journal of experimental psychology*, *20*(3), 273–281.

Wason, P. C., & Evans, J. S. B. (1974). Dual processes in reasoning? *Cognition*, *3*(2), 141–154.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824–24837.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica*, *41*(1), 67–85.