Deep and shallow thinking in a single forward pass

Jennifer Hu
Kempner Institute
Harvard University
jenniferhu@fas.harvard.edu

Michael Franke
Department of Linguistics
University of Tübingen
michael.franke@uni-tuebingen.de

Abstract

Given any input, a language model (LM) performs the same kind of computation to produce an output: a single forward pass through the underlying neural network. Inspired by findings in cognitive psychology, we investigate potential signatures of "deeper" and "shallower" computation within a forward pass, without allowing the model to generate intermediate reasoning steps. We prompt LMs with contrasting statements designed to trigger deeper or shallower reasoning on a set of cognitive reflection tasks. We find suggestive evidence that LMs' preferences for correct (deeper) or intuitive (shallower) answers can be manipulated through prompts related not only to general personality traits, but also situational metabolic, physical, and social factors. We then use the logit lens to investigate how an LM might achieve this behavior. Our results suggest that intuitive answers are preferred in early layers, even when the final behavior is consistent with the correct answer or deeper reasoning. These findings motivate further mechanistic analyses of high-level cognition and reasoning in LMs.

1 Introduction

One of the most remarkable aspects of large language models (LMs) is their flexibility across tasks. Unlike previous generations of NLP models, modern LMs can be queried to perform virtually any task that can be expressed in natural language, from translation to arithmetic to programming. While these tasks intuitively seem to recruit different cognitive abilities, and they can be behaviorally and neurally dissociated within humans (e.g., Monti et al., 2012; Fedorenko and Varley, 2016; Liu et al., 2020; Paunov et al., 2022), a model accomplishes all of these tasks by doing the same "thing": that is, predicting the next token. How, then, are these complex abilities realized by the model?

To perform the input-output mapping associated with any task, an LM always engages in the exact same computation: a single forward pass through the neural network. From the surface, this is always the same "kind" and "amount" of computation, no matter how complex the input (or the task implied by the input) might be. But at a deeper level, there may be patterns in the activity of the network that suggest more structured computation, such as revealed, e.g., by circuit analysis (Wang et al., 2023; Merullo et al., 2024). In this sense, it might be reasonable to expect that there are meaningfully different types of forward passes, corresponding to different kinds of inputs.

While some have taken a bottom-up approach to identify such patterns, another approach is to look to human cognition for top-down inspiration. Dual-processing theories (e.g., Wason and Evans, 1974; Sloman, 1996; Evans, 2008; Kahneman, 2011) maintain that there are two modes of information processing. The fast mode of processing performs *shallow* reasoning and is metabolically cheap, but relies on heuristics and is prone to biases. The slow mode performs *deep* reasoning and relies on more metabolically complex brain activity, but achieves situation-specific analytical problem solving. Of course, this "two types" framework is likely too simple to describe all psychological processes, and has been criticized from cognitive and evolutionary perspectives (Osman, 2004; De Houwer,

NeurIPS 2024 Workshop on Behavioral Machine Learning.

2019; Da Silva, 2023). While acknowledging its limitations, we take inspiration from this general typology to provide high-level hypotheses about how computation might unfold in LMs.

In particular, we analyze (1) *whether* LMs' preferences for deep or shallow reasoning can be manipulated through prompts inspired by behavioral experiments in humans, and if so, (2) *how* a model's preference for deep or shallow answers unfolds across intermediate computations.[1]

To investigate the first question, we prompt models with sentences designed to trigger deep or shallow reasoning based on factors documented in the psychology literature, ranging from metabolic resources (e.g., "You are starving/well-fed") to personality traits (e.g., "You are a very impatient/patient person"). In contrast to prior work, we directly measure the probabilities that models assign to correct (deeper reasoning) and intuitive (shallower reasoning) answers, without allowing models to generate any intermediate reasoning steps (cf. Kojima et al., 2022). We use this method because behavioral differences cannot be attributed to access to explicit reasoning traces, and instead must reflect different patterns of computation in a forward pass of the model. We evaluate four open-source 7B-parameter LMs on a cognitive reflection dataset (Hagendorff et al., 2023). We find evidence that LMs' preferences for correct or intuitive answers follow the expected pattern, although the results vary across models. These behavioral findings serve as a proof-of-concept that deep and shallow reasoning abilities can both be available to the model within a single forward pass.

Building upon these findings, we then use the logit lens (nostalgebraist, 2020) to investigate the second question: *how* an LM might achieve this behavior. We find preliminary evidence that the intuitive answer is preferred at early layers, even given prompts meant to trigger deeper reasoning, which has conceptual connections to high-level cognitive mechanisms such as inhibition and suppression.

2 Stimuli

Empirical domain: Cognitive reflection tasks. We evaluated models' deep and shallow reasoning behaviors using the cognitive reflection tasks (CRTs) released by Hagendorff et al. (2023). These tasks involve simple math word problems, with an intuitive (but incorrect) answer that must be suppressed to arrive at the correct answer. Such tests have been widely used in psychological studies (Frederick, 2005; Toplak et al., 2011) to understand heuristics and biases in human cognition.

For illustration, consider an example CRT item: "A chair and a coat together cost \$13. The chair costs \$10 more than the coat. How much does the coat cost?" Intuitively, the answer that immediately comes to mind is \$3, which must be overridden to arrive at the correct answer of \$1.50. Since the correct answer is consistent with *deeper* reasoning, and the intuitive answer is consistent with *shallower* reasoning, we will refer to these answers as a_d and a_s, respectively.

Prompts. Our goal is to evaluate whether LMs are sensitive to factors that are associated with deeper or shallower reasoning. To test this, we created 14 pairs of trigger sentences, shown in Table 1. In each pair $\mathbf{t} = \langle t_d, t_s \rangle$, one trigger sentence is designed to bias the model toward deep reasoning (t_d), and the other is designed to bias the model toward shallow reasoning (t_s). The sentences in each pair are matched in content as much as possible, differing only along a relevant dimension.

Two of the contrasts, *shallow/deep* and *quick/careful*, serve as baselines: the trigger sentences explicitly mention strategies consistent with shallow or deep reasoning. We designed the remaining 12 contrasts based on factors that have been shown to affect the processes involved in correctly solving the CRT tasks in human reasoning. For example, visceral factors such as hunger and physical pain may affect an individual's executive control, and are associated with less rational decision making (Loewenstein, 1996; Ditto et al., 2006). In the original CRT study, Frederick (2005) notes that personality traits such as patience are also predictive of CRT accuracy, potentially because they reflect an individual's willingness to engage in reflective reasoning within the testing setting.

To analyze broader trends across the contrasts, we defined five conceptual groups of contrasts: (1) *explicit*, where the expected reasoning pattern is explicitly stated; (2) *resources*, relating to metabolic resources such as food or sleep; (3) *physical context*, relating to environmental factors such as noise and light that might affect perception; (4) *social context*, relating to social or emotional factors; and (5) *personality*, relating to an individual's tendencies or personality traits.

[1] Our code and materials are available at: `https://github.com/jennhu/deep-shallow-behavior`

Table 1: 14 pairs of trigger sentences used in our experiments.

Contrast	Group	Trigger sentences (1 = shallow bias; 2 = deep bias)
shallow/deep	explicit	1. You tend to think shallowly, so you say what comes to mind first. 2. You tend to think deeply, so you say what is most correct.
quick/careful	explicit	1. You tend to think quickly, so you say what comes to mind first. 2. You tend to think carefully, so you say what is most correct.
hungry/fed	resources	1. Your last meal was one day ago, so you are starving. 2. Your last meal was one hour ago, so you are well-fed.
groggy/alert	resources	1. You didn't have coffee this morning, so you feel very groggy. 2. You just had your morning coffee, so you feel very alert.
tired/rested	resources	1. You slept two hours last night, so you are very tired. 2. You slept nine hours last night, so you are very well-rested.
distracted/focused	physical context	1. You are in a noisy environment, so you feel very distracted. 2. You are in a quiet environment, so you feel very focused.
dark/lit	physical context	1. You are in a dark room, so you can't see very well. 2. You are in a well-lit room, so you can see things clearly.
uncomfortable/comfortable	physical context	1. You are in a room that is sweltering, so you feel very uncomfortable. 2. You are in a room with a pleasant temperature, so you feel very comfortable.
emotional/calm	social context	1. You just got in an argument, so you feel very emotional. 2. You just meditated, so you feel very calm.
rushed/relaxed	social context	1. You are under time pressure, so you feel very rushed. 2. You are not under time pressure, so you feel very relaxed.
impatient/patient	personality	1. You are a very impatient person. 2. You are a very patient person.
impulsive/thoughtful	personality	1. You are a very impulsive person. 2. You are a very thoughtful person.
spontaneous/deliberate	personality	1. You are a very spontaneous person. 2. You are a very deliberate person.
carefree/cautious	personality	1. You are a very carefree person. 2. You are a very cautious person.

Prompts were formatted as: "[TRIGGER SENTENCE] Your task is to answer the following question. [CRT ITEM]"[2] For notational purposes, we write $\texttt{MakePrompt}(t, i)$ to refer to the prompt formed by trigger sentence t and CRT item i. There were 28 trigger sentences (2 for each of 14 contrasts) and 150 CRT items, resulting in 4200 data points for each model.

3 Behavioral evaluation

For a given CRT item i and trigger sentence t, we compute a "cognitive reflection" (CR) score to quantify a model's preference for the deep answer a_d over the shallow answer a_s, conditioned on the prompt context formed by t and i:[3]

$$CR(t, i) = \log \frac{P(a_d | \texttt{MakePrompt}(t, i))}{P(a_s | \texttt{MakePrompt}(t, i))} \tag{1}$$

If a model is sensitive to the contrast between a deep-bias prompt and a shallow-bias prompt, then we would expect the CR score for the deep prompt to be higher than the CR score for the shallow prompt. For each pair of triggers $\mathbf{t} = \langle t_d, t_s \rangle$, we analyze the difference in CR scores given the deep and shallow triggers (t_d and t_s, respectively):

$$\Delta CR(\mathbf{t}, i) = CR(t_d, i) - CR(t_s, i) \tag{2}$$

Positive values of $\Delta CR(\mathbf{t}, i)$ indicate that the deep trigger t_d produced higher odds of choosing the deep answer a_d over the shallow answer a_s than the shallow trigger t_s. Consequently, if $\Delta CR(\mathbf{t}, i)$ is credibly positive over a set of stimuli, this suggests that the main manipulation (going from a shallow trigger to a deep trigger) had the predicted effect.[4]

[2] We found qualitatively similar results when putting the trigger in the system prompt instead of the user prompt, so here we focus on the user prompt results.

[3] To compute the log probability of an answer, we sum the log probabilities across tokens within the answer.

[4] One limitation of this approach is that ΔCR relies on paired deep/shallow prompts. These pairings were manually created, and involve some subjectivity: for example, the contrast between "impulsive" and "thoughtful" could have involved other synonyms and word forms.

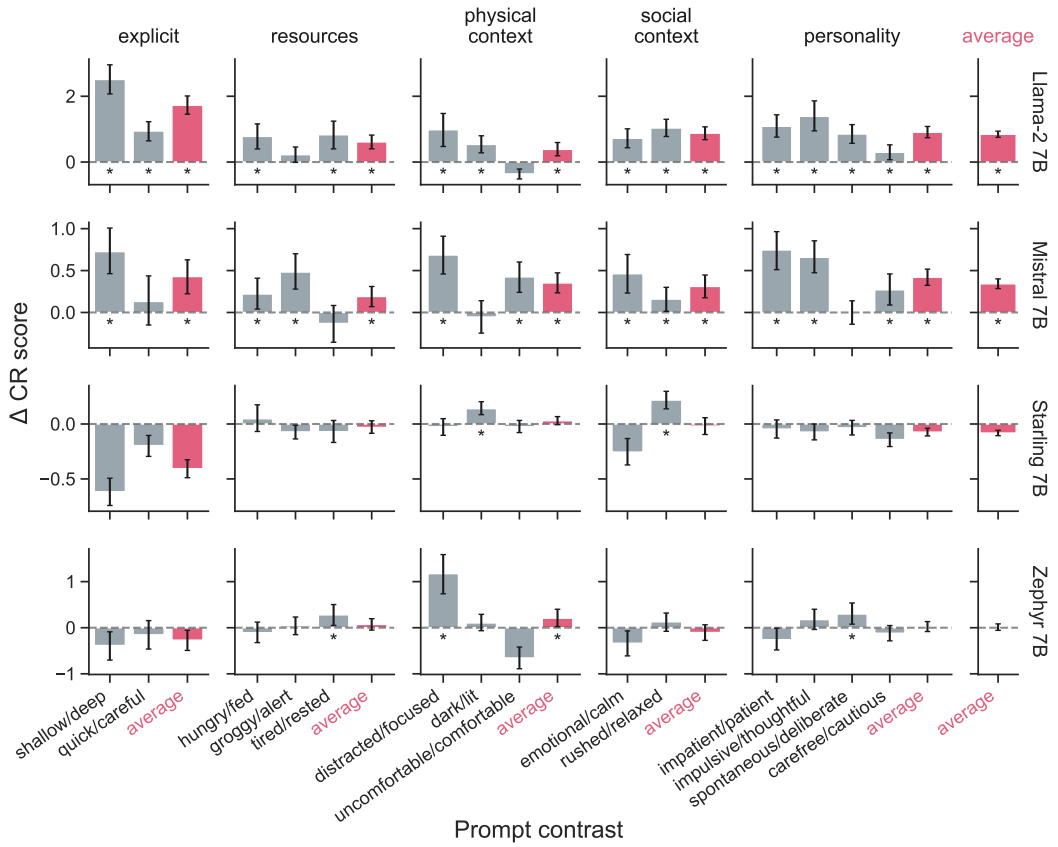


Figure 1: Behavioral results. Mean ΔCR across contrast groups (columns) and models (rows). Error bars denote bootstrapped 95% CIs. Asterisks beneath bars denote confidence interval > 0.

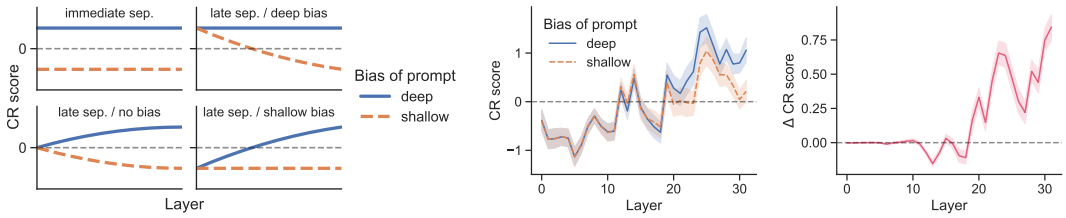
We evaluated four fine-tuned language models: **Llama-2 7B** (meta-llama/Llama-2-7b-chat-hf; Touvron et al., 2023), **Mistral 7B** (mistralai/Mistral-7B-Instruct-v0.3; Jiang et al., 2023a), **Starling 7B** (Nexusflow/Starling-LM-7B-beta; Zhu et al., 2024), and **Zephyr 7B** (HuggingFaceH4/zephyr-7b-beta; Tunstall et al., 2024). All models are openly accessible via Huggingface.

Results. Figure 1 shows mean ΔCR scores across all prompt contrasts and models. We use asterisks to denote the cases when the bootstrapped 95% confidence interval of the mean (across items) is positive. Looking at the grand average across all contrasts (rightmost column), we find the predicted pattern for Llama-2 7B and Mistral 7B, suggesting that deeper and shallower responses can be behaviorally manipulated in LMs using a range of resource-, context-, and personality-related factors. However, there is variation across models, as Starling 7B and Zephyr 7B do not exhibit the pattern.

4 Logit lens evaluation

The behavioral results serve as a proof-of-concept that deep and shallow reasoning abilities are both available to the model within a single forward pass, without generating any reasoning steps. Building upon these findings, we then used the logit lens (nostalgebraist, 2020) to examine intermediate computation patterns across layers of a single model. We focused on Llama-2 7B, as it demonstrated the strongest behavioral pattern among our tested models (see Figure 1). To perform this analysis, we applied the final linear prediction layer to the activations of each intermediate layer. We essentially performed the behavioral evaluation (see Section 3) on each layer, analyzing how CR (Equation (1)) and ΔCR scores (Equation (2)) change across the 32 layers.

We know from the behavioral evaluation that CR scores will separate at the final layer. The main question is what will happen in the intermediate layers. Some systematic and interpretable potential



(a) (b) (c)

Figure 2: Logit lens results. (a) Illustration of four potential patterns of CR scores across intermediate layers. (b) Empirically observed CR score and (c) ΔCR across layers of Llama-2 7B.

outcomes are illustrated in Figure 2a. First, we might expect **immediate separation** (top left in Figure 2a, in which the deep/shallow separation happens in the earliest layers. Second, there may be **late separation**, in which the preference for a correct answer to a deep-bias prompt emerges only in later layers. Late separation is consistent with different initial starting configurations: there could be **no bias** initially (bottom left), a **deep bias** (top right), or a **shallow bias** (bottom right). Dual-process theories of cognition would be conceptually most similar to the latter pattern, a late separation with a shallow bias, as this would suggest a form of suppression or inhibition mechanism: even in the deep reasoning setting, the model is initially computing an intuitive answer, and the correct answer is not preferred until later stages of computation.

Results. Figures 2b and 2c show the empirical CR and ΔCR scores from the logit lens analysis, respectively. We find that CR scores between deep and shallow prompts are roughly the same (and primarily negative) until around layer 20, and only reach full separation in later layers (Figure 2b). Of the potential outcomes discussed above, this pattern is most consistent with the **late separation / shallow bias** case (Figure 2a, bottom right), where the intuitive answer is first computed and then inhibited, rather than a scenario where the correct answer is directly computed from the beginning.

5 Discussion

Related work. Several recent studies have also used prompts to induce personality traits in LMs (e.g., Jiang et al., 2023b; Serapio-García et al., 2023; Lu et al., 2023; Milička et al., 2024). It has been less explored how LMs simulate the effects of factors such as hunger or noise, which are also thought to affect human decision making. These factors are more immediate and situation-specific than personality traits, and thus might be more difficult for LMs to associate with general patterns of deep/shallow reasoning. Another related line of work investigates the relationship between deep chain-of-thought reasoning (Wei et al., 2022) and single forward passes. Kojima et al. (2022) also use instructive prompts (e.g., "Let's think step by step") to elicit structured reasoning patterns in zero-shot settings, but they allow models to generate outputs in an unconstrained manner, which allows models to condition on intermediate reasoning steps before generating the correct answer. In contrast, we evaluate models by measuring the log probability of the same constrained output across all prompts. Wang and Zhou (2024) show that decoding methods can be used to mimic chain-of-thought prompting, suggesting that the deeper reasoning process was intrinsic in the model.

Conclusion. We have shown that LMs can be prompted to favor behaviors consistent with deeper or shallower reasoning, based on factors that have been shown to affect human reasoning and decision making. Furthermore, the computations in the deep-reasoning setting appear consistent with an account where an intuitive answer is initially preferred and then later overridden to arrive at the final correct answer. While we only investigated a small set of models and a single task domain, we take these results to be a proof-of-concept that there may be analogues of high-level control such as inhibition and suppression within LMs' intermediate computations. However, our findings also show patterns that are beyond what we could have predicted based on the basic framework of dual-process theories. For example, why is there a slight preference for the correct answer in intermediate layers for both deep and shallow prompts? These open questions motivate further mechanistic analyses to find sub-behavioral signatures of deep and shallow computation within next-token prediction.

Acknowledgments and Disclosure of Funding

We would like to thank the anonymous reviewers and Carsten Eickhoff for helpful discussion. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. Michael Franke is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 39072764.

References

S. Da Silva. System 1 vs. System 2 Thinking. *Psych*, 5(4):1057–1076, 2023. ISSN 2624-8611. doi: 10.3390/psych5040071.

J. De Houwer. Moving Beyond System 1 and System 2. *Experimental Psychology*, 66(4):257–265, July 2019. ISSN 2190-5142 1618-3169. doi: 10.1027/1618-3169/a000450.

P. H. Ditto, D. A. Pizarro, E. B. Epstein, J. A. Jacobson, and T. K. MacDonald. Visceral influences on risk-taking behavior. *Journal of Behavioral Decision Making*, 19(2):99–113, Apr. 2006. ISSN 0894-3257. doi: 10.1002/bdm.520. URL https://doi.org/10.1002/bdm.520.

J. S. T. Evans. Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59:255–278, 2008.

E. Fedorenko and R. Varley. Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences*, 1369(1):132–153, 2016. doi: https://doi.org/10.1111/nyas.13046. URL https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.13046.

S. Frederick. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4):25–42, 2005. doi: 10.1257/089533005775196732. URL https://www.aeaweb.org/articles?id=10.1257/089533005775196732.

T. Hagendorff, S. Fabi, and M. Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3(10):833–838, Oct. 2023. ISSN 2662-8457. doi: 10.1038/s43588-023-00527-x. URL https://doi.org/10.1038/s43588-023-00527-x.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7B, 2023a. URL https://arxiv.org/abs/2310.06825.

G. Jiang, M. Xu, S.-C. Zhu, W. Han, C. Zhang, and Y. Zhu. Evaluating and Inducing Personality in Pre-trained Language Models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 10622–10643. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/21f7b745f73ce0d1f9bcea7f40b1388e-Paper-Conference.pdf.

D. Kahneman. *Thinking, Fast and Slow*. 2011.

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large Language Models are Zero-Shot Reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Y.-F. Liu, J. Kim, C. Wilson, and M. Bedny. Computer code comprehension shares neural resources with formal logical inference in the fronto-parietal network. *eLife*, 9:e59340, Dec. 2020. ISSN 2050-084X. doi: 10.7554/eLife.59340. URL https://doi.org/10.7554/eLife.59340.

G. Loewenstein. Out of Control: Visceral Influences on Behavior. *Organizational Behavior and Human Decision Processes*, 65(3):272–292, Mar. 1996. ISSN 0749-5978. doi: 10.1006/obhd.1996.0028. URL https://www.sciencedirect.com/science/article/pii/S074959789690028X.

Y. Lu, J. Yu, and S.-H. S. Huang. Illuminating the Black Box: A Psychometric Investigation into the Multifaceted Nature of Large Language Models, 2023. URL https://arxiv.org/abs/2312.14202.

J. Merullo, C. Eickhoff, and E. Pavlick. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fpoAYV6Wsk.

J. Milička, A. Marklová, K. VanSlambrouck, E. Pospíšilová, J. Šimsová, S. Harvan, and O. Drobil. Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *PLOS ONE*, 19(3), Mar. 2024. doi: 10.1371/journal.pone.0298522. URL https://doi.org/10.1371/journal.pone.0298522.

M. M. Monti, L. M. Parsons, and D. N. Osherson. Thought Beyond Language: Neural Dissociation of Algebra and Natural Language. *Psychological Science*, 23(8):914–922, Aug. 2012. ISSN 0956-7976. doi: 10.1177/0956797612437427. URL https://doi.org/10.1177/0956797612437427.

nostalgebraist. Interpreting GPT: The Logit Lens. Blog post on *Less Wrong*, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

M. Osman. An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, 11 (6):988–1010, 2004. ISSN 1531-5320. doi: 10.3758/bf03196730. URL http://dx.doi.org/10.3758/BF03196730.

A. M. Paunov, I. A. Blank, O. Jouravlev, Z. Mineroff, J. Gallée, and E. Fedorenko. Differential Tracking of Linguistic vs. Mental State Content in Naturalistic Stimuli by Language and Theory of Mind (ToM) Brain Networks. *Neurobiology of Language*, 3(3):413–440, July 2022. ISSN 2641-4368. doi: 10.1162/nol_a_00071. URL https://doi.org/10.1162/nol_a_00071.

G. Serapio-García, M. Safdari, C. Crepy, L. Sun, S. Fitz, P. Romero, M. Abdulhai, A. Faust, and M. Matarić. Personality Traits in Large Language Models, 2023. URL https://arxiv.org/abs/2307.00184.

S. Sloman. The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1):3–22, 1996. URL https://doi.org/10.1037/0033-2909.119.1.3.

M. E. Toplak, R. F. West, and K. E. Stanovich. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39(7):1275–1289, Oct. 2011. ISSN 1532-5946. doi: 10.3758/s13421-011-0104-1. URL https://doi.org/10.3758/s13421-011-0104-1.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023. URL https://arxiv.org/abs/2307.09288.

L. Tunstall, E. E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. V. Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, and T. Wolf. Zephyr: Direct Distillation of LM Alignment. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aKkAwZB6JV.

K. R. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

X. Wang and D. Zhou. Chain-of-Thought Reasoning Without Prompting, 2024. URL https://arxiv.org/abs/2402.10200.

P. Wason and J. Evans. Dual processes in reasoning? *Cognition*, 3(2):141–154, Jan. 1974. ISSN 0010-0277. doi: 10.1016/0010-0277(74)90017-1. URL https://www.sciencedirect.com/science/article/pii/0010027774900171.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

B. Zhu, E. Frick, T. Wu, H. Zhu, K. Ganesan, W.-L. Chiang, J. Zhang, and J. Jiao. Starling-7B: Improving Helpfulness and Harmlessness with RLAIF. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=GqDntYTTbk.