Signatures of human-like processing in Transformer forward passes

Jennifer Hu*
Department of Cognitive Science
Johns Hopkins University
jennhu@jhu.edu

Michael A. Lepori
Department of Computer Science
Brown University
michael_lepori@brown.edu

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

Abstract

A dominant way of using AI models to study human cognition is to evaluate whether human-derived measures are predicted by a model's output: that is, the end-product of a forward pass. However, mechanistic interpretability has begun to reveal the models' internal processes, raising the question of whether models use human-like processing strategies. We investigate the relationship between real-time processing in humans and layer-time dynamics of computation in Transformers, testing 20 open-source models in 6 domains. We find that, in the cases where we would expect decision conflict in humans, models appear to initially favor a competing incorrect answer over the correct answer. We also find that dynamic measures improve prediction of human processing measures relative to static measures. Moreover, larger models do not always show more human-like processing patterns. Our work suggests a new way of using AI models as explicit models of human processing.

1 Introduction

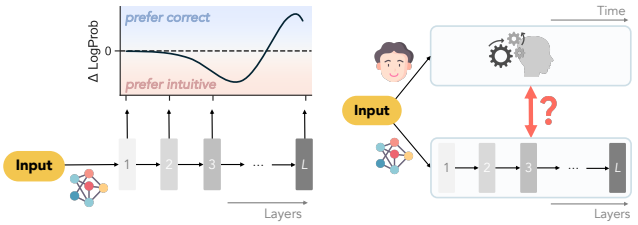
For higher-order cognitive tasks, the prevalent comparison between humans and AI models—especially language models (LMs)—is at the behavioral level. This approach typically involves using the LM to estimate the likelihoods of strings, which are linked to relevant behaviors on the human side, such as answer selections in a multiple-choice task. This "output-level" approach is often motivated on theoretical grounds, such as using probabilities derived from LMs to systematically test expectation-based theories of sentence processing (e.g., Levy, 2008; Smith and Levy, 2013).

At the same time, mechanistic interpretability has begun to uncover the internal processes that support model outputs (e.g., Biran et al., 2024; Kim et al., 2025; Wiegreffe et al., 2025). A widespread assumption is that tasks that are "easy" for an LM can be solved in fewer layer-wise computation steps (Belrose et al., 2023; Baldock et al., 2021). Recently, Kuribayashi et al. (2025) found that predictions from earlier LM layers correspond more closely with "fast" measures of human sentence processing (e.g., gaze durations), while predictions from later layers correspond with "slow" signals (e.g., the N400). Their findings raise an open question: does the *information processing* involved in a forward pass of a model resemble the *cognitive processing* involved in a human response?

Here, we investigate the relationship between layer-wise processing dynamics in Transformers and real-time processing in humans (Figure 1). We address three major research questions (RQs): (1)

*Work done while at Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University.

First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models (NeurIPS 2025).



(a) (b)

Figure 1: **Overview of our study.** (a) Exp. 1 explores whether forward passes show mechanistic signatures of competitor interference. (b) Exp. 2 investigates the ability of dynamic measures derived from forward passes to predict indicators of processing load in humans.

Do models show signs of **competitor interference effects**, with **delayed decision-making** and **two-stage processing**? (2) Do measures characterizing **(a) competitor interference effects** or **(b) other aspects of processing difficulty** in models increase the accuracy of a (linear) model **predicting human processing load**, above and beyond static measures derived from model outputs? and (3) How does **model size** affect the similarity between model and human processing?

We investigate these RQs across 20 open-source models and 6 domains, covering multiple modalities (text and vision) and human behavioral measures. We find that LM forward passes show signs of competitor interference for the items that have salient intuitive answers that compete with the ground-truth answer. Moreover, layer-time dynamics improve the ability to predict human processing indicators, above and beyond static measures derived from the final layer or an intermediate layer. Finally, we also find that larger models are not always most predictive of human processing, and mid-size LMs show the strongest signs of two-stage processing. Our results suggest that model and human processing may be facilitated or impeded by similar properties of an input stimulus.

2 Exp. 1: Competitor interference effects in LMs

One natural way to measure processing effort is to measure how the output token distribution changes throughout the layers of the model. For example, if a model is confident in the correct response to a stimulus in early layers, then that stimulus requires fairly little effort (Baldock et al., 2021). We read out token probability distributions from intermediate layers using the logit lens (nostalgebraist, 2020). Let L be the number of layers in the model; let $W_U \in \mathbb{R}^{d \times |\mathcal{V}|}$ be the unembedding matrix, where d is the hidden layer dimension and $\mathcal{V}$ is the model's vocabulary; and let $\textsc{Norm}$ be the final layer-normalization applied before submitting to the unembedding matrix. We apply the vocabulary projection W_U to the hidden representation $\mathbf{h}_{\ell,i}$ at layer $\ell \in \{1, 2, \ldots, L\}$ and token index i (conditioned on all previous tokens $t_1, t_2, \ldots, t_{i-1}$) to obtain a vector in $\mathbb{R}^{|\mathcal{V}|}$ of unnormalized logits over $\mathcal{V}$. We obtain the probability of a token t_i at the ℓ^{th} hidden layer after normalizing the logits:

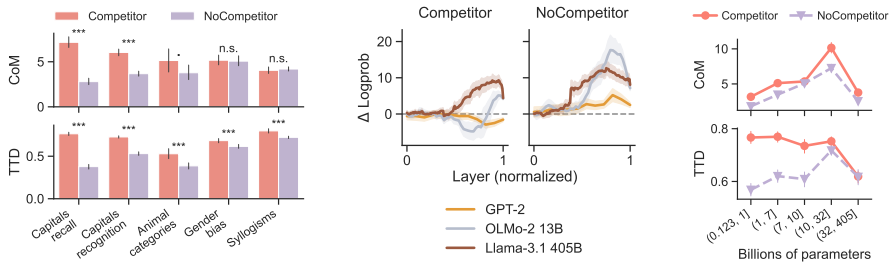
$$p(t_i \mid t_1, \ldots, t_{i-1}; \mathbf{h}_{\ell,i}) = \textsc{Softmax}\left(\textsc{Norm}(\mathbf{h}_{\ell,i})W_U\right)[\text{id}(t_i)] \quad (1)$$

To measure *relative confidence* between two tokens conditioned on a context $\mathbf{c}$ at layer $\mathbf{h}_\ell$, we define

$$\Delta\textsc{LogProb}(\mathbf{h}_\ell, v_C, v_I; \mathbf{c}) = \log p(v_C \mid \mathbf{c}; \mathbf{h}_{\ell,|\mathbf{c}|+1}) - \log p(v_I \mid \mathbf{c}; \mathbf{h}_{\ell,|\mathbf{c}|+1}) \quad (2)$$

as the log-odds ratio of the correct over the incorrect answer, where v_C and v_I refer to the first tokens of the correct and incorrect answers, respectively. We will write $\Delta\textsc{LogProb}(\mathbf{h}_\ell)$ for simplicity.

Measuring competitor interference. We define two measures that capture different aspects of competitor interference. First, we propose a novel "change of mind" measure **CoM**, which captures two-stage processing. Let m be the relative confidence given by the layer that most favors the intuitive answer over the correct answer; i.e., $m = \min_{\ell \in \{1,2,\ldots,L\}} \Delta\textsc{LogProb}(\mathbf{h}_\ell)$. We define CoM as 0 when the intuitive answer is never preferred over the correct one (i.e., when $m \geq 0$). Otherwise, CoM is given by $\min\{0, \Delta\textsc{LogProb}(\mathbf{h}_L)\} - m$. In other words, CoM is larger when there is a larger difference between m and the log-odds at the final layer $\mathbf{h}_L$. The second measure, **TtD**, captures the "time to decision": i.e., the time (in layers) at which the model begins consistently preferring the correct answer. It is defined as the last layer at which all subsequent log-odds are non-negative; i.e., $\textsc{TtD} = \max\{\ell^* \in \{1, 2, \ldots, L\} : \forall \ell \geq \ell^*, \Delta\textsc{LogProb}(\mathbf{h}_\ell) \geq 0\}$. If the model never favors the correct answer, $\textsc{TtD} = L$. We normalize TtD by L to facilitate comparison across models.



(a) (b) (c)

Figure 2: **Exp. 1 results.** (a) LMs generally show stronger signs of two-stage processing for the items with competing intuitive answers. (b) ΔLogProb across layers for sample LMs in the capitals recall domain, illustrating different processing strategies. (c) Two-stage processing interacts with size.

We considered the following empirical domains: **recall of capital cities**, **recognition of capital cities**, **animal categories** (Kieslich et al., 2020), **gender bias** (Lepori et al., 2025), and **syllogisms** (Lampinen et al., 2024). A critical subset of the test items are expected to trigger two-stage processing—for example, capital cities which are not the biggest city in its political entity (e.g., Springfield, Illinois), or categorization of atypical animal exemplars (e.g., classifying a whale as a mammal). We refer to these items as belonging to the "Competitor" condition, since they have a salient incorrect answer that intuitively "competes" with the ground-truth correct answer. See Appendix A.1 for details.

We evaluated 18 open-source, pretrained, autoregressive LMs, with 3 sizes for each of 6 families: GPT-2 (Radford et al., 2019), Llama-2 (Touvron et al., 2023), Llama-3.1 (Grattafiori et al., 2024), Gemma-2 (Gemma Team et al., 2024), OLMo-2 (Team OLMo et al., 2025), and Falcon-3 (Falcon-LLM Team, 2024). Additional details about the tested models are in Appendix A.2, Table 1.

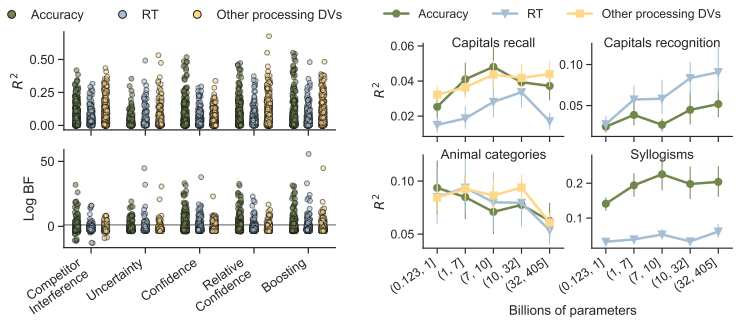
Results. LMs show signs of competitor interference: both CoM and TTD are generally higher in the Competitor condition than the NoCompetitor condition (Figure 2a).[2] To illustrate different processing strategies across conditions, Figure 2b shows ΔLogProb across layers for three example models in the capitals recall domain. We first focus on the Competitor condition (left facet). OLMo-2 13B (a mid-sized model) shows two-stage processing, while GPT-2 (the smallest model) consistently prefers the intuitive answer, and Llama-3.1 405B (the largest model) consistently prefers the correct answer. In the NoCompetitor condition (right), each model consistently prefers the correct answer. These patterns suggest interactions between model size and competitor interference. To investigate this further, we analyzed CoM and TTD across bins of model sizes, shown in Figure 2c. Mid-size models tend to show higher CoM, regardless of condition, and the difference between TTD for Competitor and NoCompetitor stimuli descreases as model size increases.

3 Exp. 2: Systematic comparison of human and model processing dynamics

Next, we investigate the ability of a larger set of measures to predict empirical measures of human processing. Full details on these metrics are given in Appendix B. On top of CoM and TTD (Section 2), the new measures are derived from five base metrics which capture different aspects of the model's decision-making at a given layer: (1) **entropy**; (2) **reciprocal rank** and (3) **log probability** of the first token of the correct answer; (4) **difference in log probability** between the correct and intuitive answers (i.e., $\Delta\text{LogProb}(\mathbf{h}_\ell)$); and (5) a **boosting** metric which characterizes how a layer changes the logit of the correct or incorrect answers.

Across these metrics, we derive two types of quantities: **static** measures, which represent the application of a metric to a single layer; and **dynamic** measures, which aggregate the metric values across all layers into a single number. Our main focus is the dynamic measures; the static measures are a baseline, representing the kinds of measures typically used to link models and human behavior.

[2]The exception is CoM in the reasoning-based domains (gender bias and syllogisms).



(a) (b)

Figure 3: **Exp. 2 results for text domains.** (a) Top: R^2 achieved by model processing measures (x-axis) across groups of human DVs (hue). Bottom: Log Bayes Factor comparing critical to baseline regression models. Horizontal line = $\log(3)$.

In each domain, we consider several task-specific empirical measures of human processing load (see Appendix A.1). We first ask whether model processing dynamics predict these human measures by computing R^2 at the item level. We then test whether the dynamic measures *improve* prediction of human measures *beyond static measures*. For each LM, model processing measure, and human DV,[3] we computed the Bayes factor between a **baseline** mixed-effects regression model (which includes the interaction between all static predictors, and random intercepts for participant) and a **critical** model (which additionally includes the model processing measure).[4] See Appendix C for details.

Here, we also considered a visual object recognition domain using out-of-distribution images (Geirhos et al., 2021). For the visual task, we evaluated two open-source vision Transformer models: ViT Small and ViT Base (Dosovitskiy et al., 2021).

Results. We now return to RQ 2(a): do the measures of competitor interference (introduced in Section 2) predict human accuracy and processing load, *above and beyond static output measures*? Each point in Figure 3a (top) shows the proportion of variance of a particular human DV explained by a particular processing metric for a given LM and domain. The competitor interference measures (left group) are often strongly correlated with human DVs. Moreover, when comparing the baseline and critical regression models (Figure 3a, bottom), many of the critical models featuring competitor interference measures (left group) improve upon the baseline models. In other words, competitor interference metrics *improve* prediction of human DVs, beyond the output measures.

Next we turn to the broader set of processing measures introduced above, addressing RQ 2(b). We again find there are many settings where these processing measures explain substantial variance in human DVs (Figure 3a, top),[5] and many of the critical models substantially improve upon the baselines (Figure 3a, bottom). These results also hold in the vision domain (Appendix D.3) and with respect to the baselines formed by static readouts from the midpoint layer (Appendix D.4).

Finally, we turn to RQ 3. Figure 3b shows the mean R^2 values achieved across groups of DVs (hue) and model sizes (x-axis) in the main LM experiments. It is *not* the case that larger models explain the most variance in human DVs. These results seem to generalize prior findings for prediction of human reading times (Oh and Schuler, 2023; Shain et al., 2024) to a larger set of empirical measurements.

4 General discussion

We found that "processing" metrics derived from forward pass dynamics predicted human task accuracy and processing measures, above and beyond static metrics derived from models' final-layer logits. In addition, larger models are not always most predictive of human processing, and mid-sized models are most likely to show competitor interference effects. An important direction for future

[3] Unless otherwise noted, we considered all trials for predicting human accuracy DVs, but restricted the analysis to trials where humans responded correctly for predicting human processing-related DVs.

[4] We also ran this analysis using static measures from an intermediate layer (i.e., the midpoint between the first and last layers) as the baseline predictors. Results from this baseline are in Appendix D.4, Figure 9.

[5] Many of these combinations result in R^2 values near 0. This is not problematic, since we don't expect *every* combination of models, model processing measures, and human measures to be strongly correlated.

work is to understand what properties of a model make it more or less human-like in its processing patterns (c.f. Wilcox et al., 2020; Oh and Schuler, 2023; Michaelov et al., 2024).

From a cognitive perspective, our work serves as a proof-of-concept demonstrating high-level alignment between human and machine processing. Our findings motivate further studies that use model processing dynamics to generate new predictions about human processing difficulty for novel task settings or stimuli.

Acknowledgments and Disclosure of Funding

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. MF is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 39072764.

References

R. Baldock, H. Maennel, and B. Neyshabur. Deep learning through the lens of example difficulty. *Advances in Neural Information Processing Systems*, 34:10876–10889, 2021.

N. Belrose, Z. Furman, L. Smith, D. Halawi, I. Ostrovsky, L. McKinney, S. Biderman, and J. Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens, 2023. URL https://arxiv.org/abs/2303.08112. _eprint: 2303.08112.

E. Biran, D. Gottesman, S. Yang, M. Geva, and A. Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14113–14130, 2024.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Falcon-LLM Team. The Falcon 3 Family of Open Models, Dec. 2024. URL https://huggingface.co/blog/falcon3.

J. F. Fiotto-Kaufman, A. R. Loftus, E. Todd, J. Brinkmann, K. Pal, D. Troitskii, M. Ripa, A. Belfki, C. Rager, C. Juang, A. Mueller, S. Marks, A. S. Sharma, F. Lucchetti, N. Prakash, C. E. Brodley, A. Guha, J. Bell, B. C. Wallace, and D. Bau. NNsight and NDIF: Democratizing Access to Open-Weight Foundation Model Internals. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=MxbEiFRf39.

R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, and W. Brendel. Partial success in closing the gap between human and machine vision. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=QkljT4mrfs.

Gemma Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, J. Ferret, P. Liu, P. Tafti, A. Friesen, M. Casbon, S. Ramos, R. Kumar, C. L. Lan, S. Jerome, A. Tsitsulin, N. Vieillard, P. Stanczyk, S. Girgin, N. Momchev, M. Hoffman, S. Thakoor, J.-B. Grill, B. Neyshabur, O. Bachem, A. Walton, A. Severyn, A. Parrish, A. Ahmad, A. Hutchison, A. Abdagic, A. Carl, A. Shen, A. Brock, A. Coenen, A. Laforge, A. Paterson, B. Bastian, B. Piot, B. Wu, B. Royal, C. Chen, C. Kumar, C. Perry, C. Welty, C. A. Choquette-Choo, D. Sinopalnikov, D. Weinberger, D. Vijaykumar, D. Rogozińska, D. Herbison, E. Bandy, E. Wang, E. Noland, E. Moreira, E. Senter, E. Eltyshev, F. Visin, G. Rasskin, G. Wei, G. Cameron, G. Martins, H. Hashemi, H. Klimczak-Plucińska, H. Batra, H. Dhand, I. Nardini, J. Mein, J. Zhou, J. Svensson, J. Stanway, J. Chan, J. P. Zhou, J. Carrasqueira, J. Iljazi, J. Becker, J. Fernandez, J. v. Amersfoort, J. Gordon, J. Lipschultz, J. Newlan, J.-y. Ji, K. Mohamed, K. Badola, K. Black,

K. Millican, K. McDonell, K. Nguyen, K. Sodhia, K. Greene, L. L. Sjoesund, L. Usui, L. Sifre, L. Heuermann, L. Lago, L. McNealus, L. B. Soares, L. Kilpatrick, L. Dixon, L. Martins, M. Reid, M. Singh, M. Iverson, M. Görner, M. Velloso, M. Wirth, M. Davidow, M. Miller, M. Rahtz, M. Watson, M. Risdal, M. Kazemi, M. Moynihan, M. Zhang, M. Kahng, M. Park, M. Rahman, M. Khatwani, N. Dao, N. Bardoliwalla, N. Devanathan, N. Dumai, N. Chauhan, O. Wahltinez, P. Botarda, P. Barnes, P. Barham, P. Michel, P. Jin, P. Georgiev, P. Culliton, P. Kuppala, R. Comanescu, R. Merhej, R. Jana, R. A. Rokni, R. Agarwal, R. Mullins, S. Saadat, S. M. Carthy, S. Cogan, S. Perrin, S. M. R. Arnold, S. Krause, S. Dai, S. Garg, S. Sheth, S. Ronstrom, S. Chan, T. Jordan, T. Yu, T. Eccles, T. Hennigan, T. Kocisky, T. Doshi, V. Jain, V. Yadav, V. Meshram, V. Dharmadhikari, W. Barkley, W. Wei, W. Ye, W. Han, W. Kwon, X. Xu, Z. Shen, Z. Gong, Z. Wei, V. Cotruta, P. Kirk, A. Rao, M. Giang, L. Peran, T. Warkentin, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, D. Sculley, J. Banks, A. Dragan, S. Petrov, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, S. Borgeaud, N. Fiedel, A. Joulin, K. Kenealy, R. Dadashi, and A. Andreev. Gemma 2: Improving Open Language Models at a Practical Size, 2024. URL https://arxiv.org/abs/2408.00118.

D. G. Goldstein and G. Gigerenzer. Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109(1):75–90, 2002.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. v. d. Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. v. d. Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. d. Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan,

I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, and Z. Ma. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783.

P. J. Kieslich, M. Schoemann, T. Grage, J. Hepp, and S. Scherbaum. Design factors in mouse-tracking: What makes a difference? *Behavior Research Methods*, 52(1):317–341, Feb. 2020. ISSN 1554-3528. doi: 10.3758/s13428-019-01228-y. URL https://doi.org/10.3758/s13428-019-01228-y.

G. Kim, M. Valentino, and A. Freitas. A Mechanistic Interpretation of Syllogistic Reasoning in Auto-Regressive Language Models, 2025. URL https://arxiv.org/abs/2408.08590. _eprint: 2408.08590.

T. Kuribayashi, Y. Oseki, S. B. Taieb, K. Inui, and T. Baldwin. Large language models are human-like internally, 2025. URL https://arxiv.org/abs/2502.01615.

A. K. Lampinen, I. Dasgupta, S. C. Y. Chan, H. R. Sheahan, A. Creswell, D. Kumaran, J. L. McClelland, and F. Hill. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233, July 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae233. URL https://doi.org/10.1093/pnasnexus/pgae233.

M. A. Lepori, M. C. Mozer, and A. Ghandeharioun. Racing Thoughts: Explaining Large Language Model Contextualization Errors. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025. URL https://arxiv.org/abs/2410.02102.

R. Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126 – 1177, 2008. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2007.05.006. URL http://www.sciencedirect.com/science/article/pii/S0010027707001436.

J. Merullo, C. Eickhoff, and E. Pavlick. Language Models Implement Simple Word2Vec-style Vector Arithmetic. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5030–5047, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.281. URL https://aclanthology.org/2024.naacl-long.281.

J. Michaelov, C. Arnett, and B. Bergen. Revenge of the Fallen? Recurrent Models Match Transformers at Predicting Human Language Comprehension Metrics. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=amhPBLFYWv.

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

B.-D. Oh and W. Schuler. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, 11:336–350, Mar. 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00548. URL https://doi.org/10.1162/tacl_a_00548.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language Models are Unsupervised Multitask Learners, 2019. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

C. Shain, C. Meister, T. Pimentel, R. Cotterell, and R. Levy. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121, Mar. 2024. doi: 10.1073/pnas.2307876121. URL https://doi.org/10.1073/pnas.2307876121. Publisher: Proceedings of the National Academy of Sciences.

N. J. Smith and R. Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302 – 319, 2013. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2013.02.013. URL http://www.sciencedirect.com/science/article/pii/S0010027713000413.

Team OLMo, P. Walsh, L. Soldaini, D. Groeneveld, K. Lo, S. Arora, A. Bhagia, Y. Gu, S. Huang, M. Jordan, N. Lambert, D. Schwenk, O. Tafjord, T. Anderson, D. Atkinson, F. Brahman, C. Clark, P. Dasigi, N. Dziri, M. Guerquin, H. Ivison, P. W. Koh, J. Liu, S. Malik, W. Merrill, L. J. V. Miranda, J. Morrison, T. Murray, C. Nam, V. Pyatkin, A. Rangapur, M. Schmitz, S. Skjonsberg, D. Wadden, C. Wilhelm, M. Wilson, L. Zettlemoyer, A. Farhadi, N. A. Smith, and H. Hajishirzi. 2 OLMo 2 Furious, 2025. URL https://arxiv.org/abs/2501.00656.

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.

A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232, 1973. doi: 10.1016/0010-0285(73)90033-9. URL http://dx.doi.org/10.1016/0010-0285(73)90033-9.

M. G. Vilas, T. Schaumlöffel, and G. Roig. Analyzing vision transformers for image classification in class embedding space. *Advances in neural information processing systems*, 36:40030–40041, 2023.

S. Wiegreffe, O. Tafjord, Y. Belinkov, H. Hajishirzi, and A. Sabharwal. Answer, Assemble, Ace: Understanding How LMs Answer Multiple Choice Questions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6NNA0MxhCH.

E. Wilcox, J. Gauthier, J. Hu, P. Qian, and R. Levy. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the Cognitive Science Society*, 2020. URL https://arxiv.org/abs/2006.01912.

A Details of experiments

A.1 Empirical domains & human experiments

Below, we provide additional details about the empirical domains and how the human data were collected, processed, and analyzed. The tasks are illustrated in Figure 4.

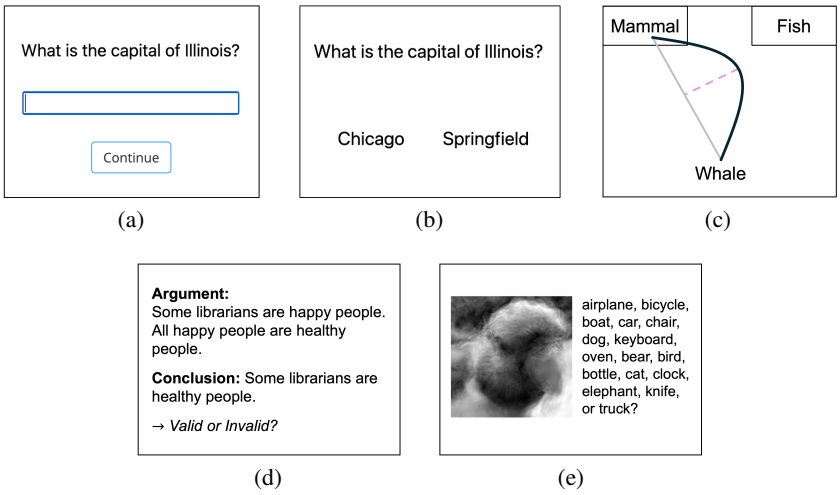


Figure 4: Illustration of human tasks analyzed in Experiment 2. (a) Recall (free response) of capital cities. (b) Recognition (forced-choice) of capital cities. (c) Categorization of typical and atypical animal exemplars via mouse movement (Kieslich et al., 2020). (d) Judgment of logical validity of syllogistic arguments (Lampinen et al., 2024). (e) Object recognition of out-of-distribution images (Geirhos et al., 2021).

Capitals recall. We begin with a standard fact recall task: retrieving the capital city of a country or United States state (Merullo et al., 2024). We manually curated 62 items (22 countries, 40 states), each consisting of a political entity, the correct capital city, and an incorrect city in the entity. The Competitor condition contained 42 items where the incorrect answer is the most populous city within its political entity (e.g., Illinois/Springfield/Chicago), which may potentially trigger competitor interference effects via reasoning similar to availability (Tversky and Kahneman, 1973) or recognition heuristics (Goldstein and Gigerenzer, 2002). The NoCompetitor condition contained 20 items, where the incorrect answer is not expected to compete with the correct answer (e.g., France/Paris/Marseilles).

We recruited 56 participants on Prolific, based in the United States with a self-reported native language of English. Participants were paid at a rate of $8/hr. Each participant saw each of the 42 Competitor items in randomized order, one on each trial. On each trial, participants saw a question of the form "What is the capital of <ENTITY>?" and freely typed their answer in a text box (see Figure 4a). Each keystroke and its timestamp was logged. The text box had to be non-empty in order to advance to the next trial. Before the experiment, participants were asked to certify that they would not use any external tools such as search engines or generative AI to perform the study. They were also informed that their payment did not depend on the correctness of their answers, and were encouraged to guess if they were unsure of the answer.

For analyses, we excluded trials where the RT (between stimulus onset and submission of data) is more than 2 standard deviations away from the mean RT across all participants' trials. We also excluded trials (~5%) where the total number of keystrokes was less than the number of characters in the final answer. This occurred occasionally when people copied and pasted their answers, or when people's keystrokes were not recorded due to browser settings.

We consider 6 human behavioral DVs. First, we consider 2 measures of accuracy: **strict**, where a response is considered correct if it is an exact string match with the correct answer (after removing casing and whitespaces), and **lenient**, which allows for minor typos and spelling variations. We used GPT-4o to code responses for "lenient" accuracy (see Appendix A.3). Next, we consider 4 measures of processing load, consisting of **response time (RT)** and 3 measures derived from typing patterns: **time of the first keypress after the last time the box was empty**[6] (a measure of how long a participant "thinks" before typing their final answer), the ratio between the **total # of keypresses** and the length (in characters) of the final answer, and the **# of backspace presses** (a proxy for a participant's uncertainty).

[6]Due to technical difficulties, this particular DV was only recorded for 30 participants out of the total 56.

Capitals recognition. Using the same stimuli as above, we then tested factual knowledge of capital cities in a *recognition* (i.e., forced choice) setting, where both the correct and intuitive options are presented.

We recruited 41 participants on Prolific, based in the United States with a self-reported native language of English. Participants were paid at a rate of $10.95/hr. On each trial, participants saw a question like "What is the capital of <ENTITY>?" and two answer options (correct and intuitive) beneath it (see Figure 4b). Their task was to press the "f" key to select the answer on the left-hand side of the screen, and the "j" key to select the answer on the right-hand side. Exactly half of the trials presented the correct answer on the left side, and the other half on the right side. As in the recall experiment (see above), each participant saw each of the 42 Competitor items, and the order of trials was randomized at runtime for each participant. Again, we excluded trials with response times more than 2 standard deviations away from the mean.

Here, we only predict two human DVs of interest: **accuracy** and **RT**.

Animal categories. Categorization of atypical exemplars (e.g., categorizing a whale as a mammal) can also induce processing difficulty. We used the stimuli and human data (N=108 participants) from Experiment 1 of Kieslich et al. (2020). The stimuli consist of 19 animals, paired with a correct category and an incorrect category. There are 6 items in the Competitor condition, where the animal is an *atypical* exemplar of the correct category (e.g., whale/mammal/fish), and 13 items in the NoCompetitor condition, where the animal is a *typical* exemplar (e.g., salmon/fish/mammal).

The human data consist of sequences of triples (t, x, y): a timestamp t, and x- and y-coordinates of the participant's mouse. There are 108 participants in total: 54 in the "click" group, where participants had to click on the region of the screen displaying their chosen category; and 54 in the "touch" group, where participants merely needed to move their mouse into that region of the screen. In the regression analyses, we included the group participants were assigned to as a fixed factor.

We considered 6 human DVs: **accuracy**, and 5 indicators of processing load or decision conflict. The processing-related DVs included **RT**, and 4 measures derived from the spatial mouse trajectories: **AUC** (area between the trajectory and a straight line from the start to the selected option), **MAD** (signed maximum deviation between the trajectory and a straight line from the start to the selected option), **# x flips** (number of directional changes on the x-axis), and **time of maximum acceleration**.

Syllogisms. Next, we explore a more challenging and practically relevant task: judging the logical validity of simple syllogistic arguments. We used the stimuli and human data from Lampinen et al. (2024). The stimuli consist of 192 simple syllogistic arguments (two premises and a candidate conclusion). The correct answer is either "valid" or "invalid", depending on the ground-truth logical validity of the conclusion, and the incorrect answer is either "invalid" or "valid". The Competitor condition includes the 48 stimuli which induce *content effects*: i.e., when the logical validity of the conclusion is inconsistent with people's prior beliefs, thus triggering competition from the intuitive but incorrect answer. The remaining 144 items are in the NoCompetitor condition.

We used the stimuli and human data from Lampinen et al. (2024). Note that the data do not contain subject-level identifiers, but there are multiple variations of each item, so we include random intercepts for each item in the regression analyses. The stimuli consist of 192 items of the form $(e, \mathbf{a}^*, \mathbf{a}')$, where e is the argument (two premises) and a candidate conclusion; $\mathbf{a}^*$ is "valid" or "invalid", depending on the ground truth of whether the conclusion logically follows from the argument; and $\mathbf{a}'$ is the incorrect label ("invalid" or "valid"). There are 96 "realistic" items: 48 where the logical validity of the argument is *consistent* with prior beliefs, and 48 where the logical validity is *inconsistent* with prior beliefs, thus inducing a content effect. In addition, there are 96 "nonsense" items, which contain nonsense words in place of semantically contentful words (e.g., "Argument: Some pand are ing. All ing are phrite. Conclusion: Some pand are phrite.").

Here, we only predict two human DVs of interest: **accuracy** and **RT**.

Object recognition. Finally, we tested whether our approach would generalize to an entirely different modality: vision. We compared pre-trained vision Transformer models (ViTs) and humans on their out-of-distribution (OOD) object recognition abilities, using the stimuli and human data released by Geirhos et al. (2021). The stimuli include 17 datasets of OOD images. The images feature objects in 16 basic ImageNet categories (e.g., chair, dog), but with manipulations such as stylization

Table 1: Overview of models evaluated in our experiments. (a) Language models. (b) Vision models.

(a)

Model	HuggingFace ID	# params (B)	# layers	Vocab size	Training
GPT-2	`gpt2`	0.124	12	50K	40 GB
GPT-2 Med	`gpt2-medium`	0.355	24	50K	40 GB
GPT-2 XL	`gpt2-xl`	1.5	48	50K	40 GB
Llama-2 7B	`meta-llama/Llama-2-7b-hf`	7	32	32K	2T
Llama-2 13B	`meta-llama/Llama-2-13b-hf`	13	40	32K	2T
Llama-2 70B	`meta-llama/Llama-2-70b-hf`	70	80	32K	2T
Llama-3.1 8B	`meta-llama/Llama-3.1-8B`	8	32	128K	15T+
Llama-3.1 70B	`meta-llama/Llama-3.1-70B`	70	80	128K	15T+
Llama-3.1 405B	`meta-llama/Llama-3.1-405B`	405	126	128K	15T+
Gemma-2 2B	`google/gemma-2-2b`	2	26	256K	2T
Gemma-2 9B	`google/gemma-2-9b`	9	42	256K	8T
Gemma-2 27B	`google/gemma-2-27b`	27	46	256K	13T
OLMo-2 7B	`allenai/OLMo-2-1124-7B`	7	32	100K	4T
OLMo-2 13B	`allenai/OLMo-2-1124-13B`	13	40	100K	5T
OLMo-2 32B	`allenai/OLMo-2-0325-32B`	32	64	100K	6T
Falcon-3 1B	`tiiuae/Falcon3-1B-Base`	1	18	131K	80 GT
Falcon-3 3B	`tiiuae/Falcon3-3B-Base`	3	22	131K	100 GT
Falcon-3 10B	`tiiuae/Falcon3-10B-Base`	10	40	131K	2 TT

(b)

Model	pytorch-image-models ID	# params (B)	# layers
ViT Small	`vit-small-patch16-224`	0.022	12
ViT Base	`vit-base-patch16-224`	0.086	12

or parametric degradations. For the 12 parametric image degradation datasets, images are subject to different levels of degradation; e.g., different levels of uniform noise. We include a `condition` variable in our baseline and critical models for these 12 datasets to account for this. This factor was omitted for the 5 non-parametric manipulations.

Each item is of the form $(I, \mathbf{a}^*)$, where I is a 224×224 image, and $\mathbf{a}^*$ is the correct category out of the 16 possible options.

Analogously to the logit lens for LMs, we derive intermediate model predictions by applying the final layernorm to the representation of the classification token after every intermediate layer, followed by the classification head. Since ViTs are encoder-only models (i.e., not autoregressive), there is less training pressure to build up classification decisions in a single residual stream, which is necessary for deriving our processing metrics. Nevertheless, Vilas et al. (2023) have found that class representations are built up across ViT layers.

Since the stimuli do not have paired "incorrect" answers, we only measure uncertainty (i.e., entropy) and confidence in the correct answer (i.e., the reciprocal rank or log probability of the correct image category). Again, we only considered two human DVs: **accuracy** and **RT**.

A.2 Model evaluation

We evaluated models using the `nnsight` package (Fiotto-Kaufman et al., 2025). Table 1 provides more details about the models evaluated in our experiments. Inference was run on NVIDIA A100-40GB and H100 80GB GPUs during April–May 2025.

Below, we provide the full input for evaluating LMs in each domain.

Capitals recall. For each item, we measured model predictions conditioned on a context like "The capital of Illinois is".

Capitals recognition. For each item $(e, \mathbf{a}^*, \mathbf{a}')$, we constructed two prefix contexts, $\mathbf{c}_1$ and $\mathbf{c}_2$:

$$\mathbf{c}_1 = \text{“The capital of } e \text{ is either } \mathbf{a}^* \text{ or } \mathbf{a}'\text{. In fact, the capital of } e \text{ is”} \quad (3)$$

$$\mathbf{c}_2 = \text{“The capital of } e \text{ is either } \mathbf{a}' \text{ or } \mathbf{a}^*\text{. In fact, the capital of } e \text{ is”} \quad (4)$$

We then average all relevant metrics across these two orderings.

Animal categories. We measured each LM's responses conditioned on the prefix $\mathbf{c}$ = “A e is a type of” (or “An e is a type of”, depending on the first letter of the animal name).

We translated the materials from Kieslich et al. (2020) from German into English for evaluating LMs.

Syllogisms. We measured LM responses conditioned on the following prefix, which is slightly modified from the prompt used by Lampinen et al. (2024):

```
In this task, you will have to answer a series of questions. You will have to choose
the best answer to complete a sentence, paragraph, or question. Please answer them to the
best of your ability.

Please assume that the first two sentences in the argument are true. Determine whether
the argument is valid or invalid, that is, whether the conclusion follows from the first
two sentences:
Argument: <ARGUMENT>
Conclusion: <CONCLUSION>
Answer: The argument is
```

A.3 Annotating responses in capitals recall experiment (Study 1a)

We used the following prompt to query GPT-4o (April 2025) through the OpenAI API for labeling the responses from the capitals recall experiment.

```
People were asked to name the capital city of various countries and US states,
and your job is to label their responses. The possible labels are "Correct"
(the correct capital), "Intuitive" (the intuitive capital), "Alternate city"
(another city in the entity), "Not sure" (expressions of uncertainty), or "Other".

It's ok if the responses include minor typos or spelling variations.
Please provide the label that best describes the response.

Here are some examples.

Entity: Illinois
Correct: Springfield
Intuitive: Chicago
Response: "springfield"
Label: Correct

Entity: Pennsylvania
Correct: Harrisburg
Intuitive: Philadelphia
Response: "Philladelphia"
Label: Intuitive

Entity: Morrocco
Correct: Rabat
Intuitive: Marrakesh
Response: "Casablanca"
Label: Alternate city

Entity: Maryland
Correct: Annapolis
Intuitive: Baltimore
Response: "idk"
```

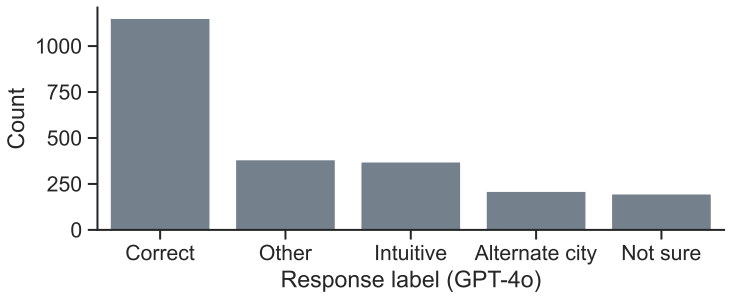


Figure 5: Distribution of labels assigned by GPT-4o to free responses in Study 1a (capitals recall).

```
Label: Not sure

Entity: Canada
Correct: Ottawa
Intuitive: Toronto
Response: "Canada"
Label: Other

Now, here is a new item for you to label.

Entity: {entity}
Correct: {correct}
Intuitive: {intuitive}
Response: "{response}"

How would you label this response? Only respond with "Correct", "Intuitive",
"Alternate city", "Not sure", or "Other" as your answer.
```

We included examples with minor typos (e.g., "Philladelphia") and capitalization variations (e.g., "springfield") in the prompt.

For each trial, the actual entity, correct answer, intuitive answer, and human response were substituted in the `{entity}`, `{correct}`, `{intuitive}`, and `{response}` placeholders, respectively.

The resulting distribution of labels is shown in Figure 5.

B Details of model-derived metrics

B.1 Uncertainty

We first consider the model's "uncertainty" at the moment of decision-making, given by the entropy of the next-token distribution given context $\mathbf{c}$:

$$\textsc{Entropy}(\mathbf{c};\mathbf{h}) = -\sum_{v\in\mathcal{V}} p(v|\mathbf{c};\mathbf{h}) \log p(v|\mathbf{c};\mathbf{h}) \tag{5}$$

Note that this measure depends only on the context, and not any particular answer option.

The output measure of uncertainty is given by the entropy of the output distribution at the final layer (**EntropyFinal**). The processing measures of uncertainty are given by the summed entropy across layers (**EntropyAUC**), and the layer index $\ell^* \in [1, L-1]$ of the largest *decrease* in entropy (**EntropyLayer**).

B.2 Confidence in correct answer

Next, we consider the model's degree of "confidence" in the correct answer. We consider two measures of confidence: (1) the log probability and (2) the reciprocal rank of the first token of the correct answer, both conditioned on the context $\mathbf{c}$.

Let $\mathbf{c}$ be the context (item) that the model is conditioned on, and let $\mathbf{a} = [a_1, a_2, \ldots, a_{|\mathbf{a}|}]$ be the answer string that we want to score, consisting of $|\mathbf{a}|$ tokens. We compute the log probability of the first token a_1 by applying log softmax to the logits (i.e., log on top of Equation (1)).

We also analyze the reciprocal rank of the first token a_1 of the answer $\mathbf{a}$ within the logits given by a particular layer:

$$\text{RANK}^{-1}(\mathbf{a}, \mathbf{c}; \mathbf{h}) = \frac{1}{\text{Rank}(\text{id}(a_1), \text{LOGITS}(\mathbf{h}|\mathbf{c}))} \tag{6}$$

where $\text{id}(a_1)$ gives the token index of token a_1. If a_1 is the top-ranked token, then this value will be 1, and if it is the bottom-ranked token, then this value will be $\frac{1}{|\mathcal{V}|}$.

The output measures of confidence are the reciprocal rank and log probability of the correct answer at the final layer (**RRankFinal** and **LogprobFinal**, respectively). There are four corresponding processing measures of confidence: the area under the curves (**RRankAUC**[7] and **LogprobAUC**), as well as the layer indices of largest increase (**RRankLayer** and **LogprobLayer**).

B.3 Confidence in correct answer, relative to intuitive answer

Next, we consider the model's confidence in the correct answer, *relative to* an alternate answer. In our experiments, this alternate answer is an intuitively salient (but incorrect) answer. We measure the relative confidence at a given layer $\mathbf{h}$ as the difference in log probability between the correct answer $\mathbf{a}^*$ and intuitive answer $\mathbf{a}'$, conditioned on the context $\mathbf{c}$:

$$\Delta\text{LOGPROB}(\mathbf{a}^*, \mathbf{a}'|\mathbf{c}; \mathbf{h}) = \text{LOGPROB}(\mathbf{a}^*|\mathbf{c}; \mathbf{h}) - \text{LOGPROB}(\mathbf{a}'|\mathbf{c}; \mathbf{h}) \tag{7}$$

The output measure of relative confidence is the log probability difference at the final layer (**ΔLogprobFinal**). For processing measures, we obtained three metrics based on the curve formed by the logprob differences. First, we computed two AUC-based metrics: the area *above* 0 (**ΔLogprobAUC+**), which measures the amount of "time" and confidence with which the model preferred the correct answer $\mathbf{a}^*$, and the area *below* 0 (**ΔLogprobAUC-**), which measures the amount of "time" and confidence with which the model preferred the intuitive answer $\mathbf{a}'$. Note that these two quantities are not redundant—they could both be high, both be low, or one could be high while the other is low. Finally, we computed the layer at which we see the largest increase in the log probability difference between the correct and intuitive answers (**ΔLogprobLayer**).

B.4 Boosting of correct answer, relative to intuitive answer

We consider signatures of a model "boosting" the correct answer, relative to an alternate intuitive answer. Measures derived from the output layer alone do not give information about "boosting" dynamics, so we only consider processing measures in this case.

In the residual stream view of a transformer (Elhage et al., 2021), the effect of any layer, $\mathbf{h}$, in a Transformer model can be summarized by the delta between the residual stream before and after that layer, $\Delta\mathbf{h}$. Notably, $\Delta\mathbf{h}$ is simply another vector of the same dimensionality of $\mathbf{h}$, so one can project it into the vocabulary space using logit lens. We wish to quantify how different layers promote a correct answer over an intuitive answer. To do so, we computed the "logit difference" for each item with correct answer $\mathbf{a}^*$, intuitive answer $\mathbf{a}'$, and context $\mathbf{c}$

$$\Delta\text{LOGIT}(\mathbf{a}^*, \mathbf{a}'|\mathbf{c}; \Delta\mathbf{h}) = \text{LOGIT}(a_1^*|\mathbf{c}; \Delta\mathbf{h}) - \text{LOGIT}(a_1'|\mathbf{c}; \Delta\mathbf{h}) \tag{8}$$

and the "term difference"

$$\Delta\text{TERM}(\mathbf{a}^*, \mathbf{a}'|\mathbf{c}; \Delta\mathbf{h}) = \left|\text{LOGIT}(a_1^*|\mathbf{c}; \Delta\mathbf{h})\right| - \left|\text{LOGIT}(a_1'|\mathbf{c}; \Delta\mathbf{h})\right| \tag{9}$$

between the first tokens of the correct and intuitive answer options. This gives us a tuple for each layer and item that describes (i) whether or not $\mathbf{h}$ increases the probability of generating $\mathbf{a}^*$ over $\mathbf{a}'$ and (ii) whether or not $\mathbf{h}$ primarily changes the log probability of $\mathbf{a}^*$ or the log probability $\mathbf{a}'$.

[7] Note that to compute **RRankAUC** we compute the area between the layer-wise RANK^{-1} curve and the lowest possible reciprocal rank $\frac{1}{|\mathcal{V}|}$ (which is extremely small in practice).

In this space, the direction of the $\langle 1, 1\rangle$ vector can be interpreted as *boosting* the correct answer relative to the intuitive answer — the layer is increasing the probability of generating $\mathbf{a}^*$ over $\mathbf{a}'$ by increasing the log probability of $\mathbf{a}^*$ (rather than decreasing the log probability of $\mathbf{a}'$). Similarly, the direction of the $\langle -1, -1\rangle$ vector can be interpreted as *boosting* the intuitive answer.[8]

For a given layer and item, we then compute the scalar projection of $\langle \Delta\textsc{Term}(\mathbf{a}^*, \mathbf{a}'|\mathbf{c}; \Delta\mathbf{h}), \Delta\textsc{Logit}(\mathbf{a}^*, \mathbf{a}'|\mathbf{c}; \Delta\mathbf{h})\rangle$ onto the $\langle 1, 1\rangle$ vector. For simplicity, we write this quantity as $S(\mathbf{a}^*, \mathbf{a}'|\mathbf{c}; \Delta\mathbf{h})$:

$$S(\mathbf{a}^*, \mathbf{a}'|\mathbf{c}; \Delta\mathbf{h}) = \frac{\langle \Delta\textsc{Term}(\mathbf{a}^*, \mathbf{a}'|\mathbf{c}; \Delta\mathbf{h}), \Delta\textsc{Logit}(\mathbf{a}^*, \mathbf{a}'|\mathbf{c}; \Delta\mathbf{h})\rangle \cdot \langle 1, 1\rangle}{\|\langle 1, 1\rangle\|} \tag{10}$$

We used the layer-wise scalar projection curve to derive three processing measures: the area *above* 0, which represents time and "strength" of boosting the correct answer (**BoostAUC+**); the area *below* 0, which represents time and "strength" of boosting the intuitive answer (**BoostAUC-**), and **BoostLayer**, the layer with the largest scalar projection. Note that here we are looking at the maximum *value* instead of the maximum *change*, as in the other metric groups, since the projection is already a measure of change (i.e., boosting).

C Details of regression analyses

For each LM and each human DV of interest, we first fit a strong **baseline** mixed-effects regression model, which includes the interaction between all static predictors derived from the final layer, and random intercepts for participant (Equation (11)). For each dynamic measure, we then fit a new **critical** model that additionally includes the new independent variable (IV) (Equation (12)). We then compute the Bayes factor between the baseline and critical regression models.

```
DV ~ staticEntropy * staticRRank * staticLogProb * staticΔLogProb + (1|Subject)
```
(11)

```
DV ~ IV + staticEntropy * staticRRank * staticLogProb * staticΔLogProb + (1|Subject)
```
(12)

To perform the model comparisons for each study, we fit (generalized) linear mixed-effects models in R (version 4.3.2) using the `lme4` package. All predictors were centered and scaled. In some cases, a predictor had only one unique value across all items (e.g., the rank of the correct answer always being 1), in which case the predictor was dropped from the analysis for that particular model and task.

We fit standard linear models using maximum likelihood for all DVs except for binary variables (accuracy) and count variables (number of backspaces, number of x-position flips), in which case we fit generalized linear mixed-effects models using `family=binomial(link='logit')` and `family='poisson'`, respectively. Time-based DVs (such as RT or max acceleration time) were log-transformed.

Log Bayes factors were computed on top of fitted `lme4` models using the `bayestestR` package.

D Additional experimental results

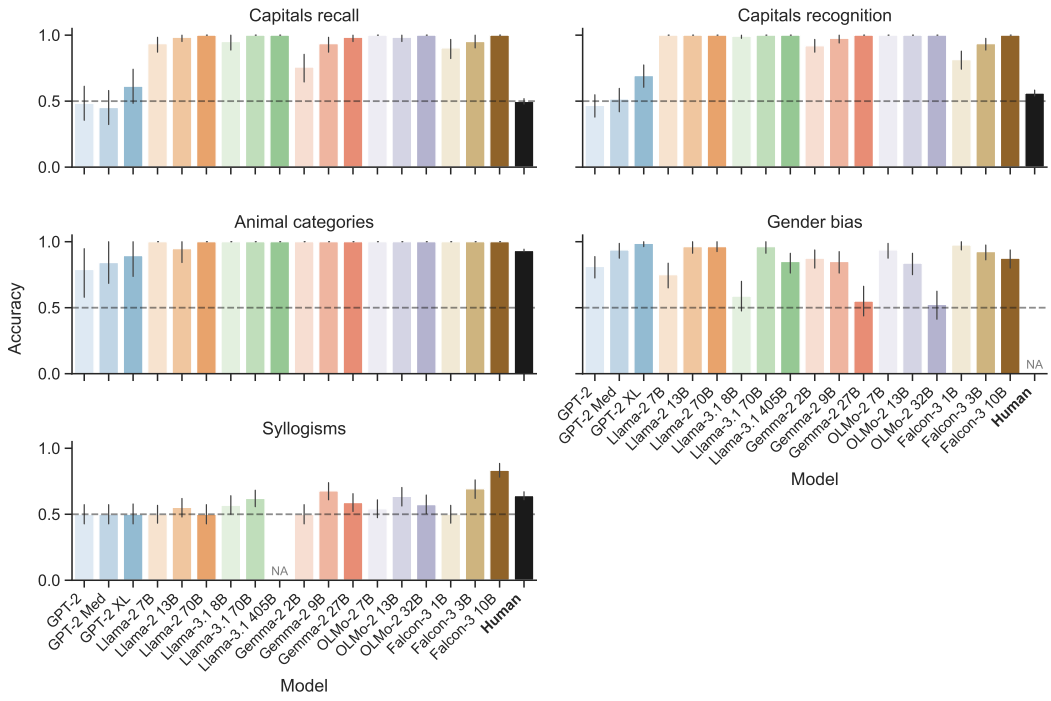
D.1 Task accuracy

Figure 6 shows the overall accuracy achieved by each model and humans on each task. For text-based tasks, model accuracy is defined as whether the model assigned higher total probability to the correct answer option (i.e., summed log probability across tokens) than the incorrect answer option. For vision-based tasks, model accuracy is defined as whether the model assigned highest probability to the correct image category (out of the 16 options). This notion of accuracy is evaluated in the "normal" way; i.e., at the final layer only.

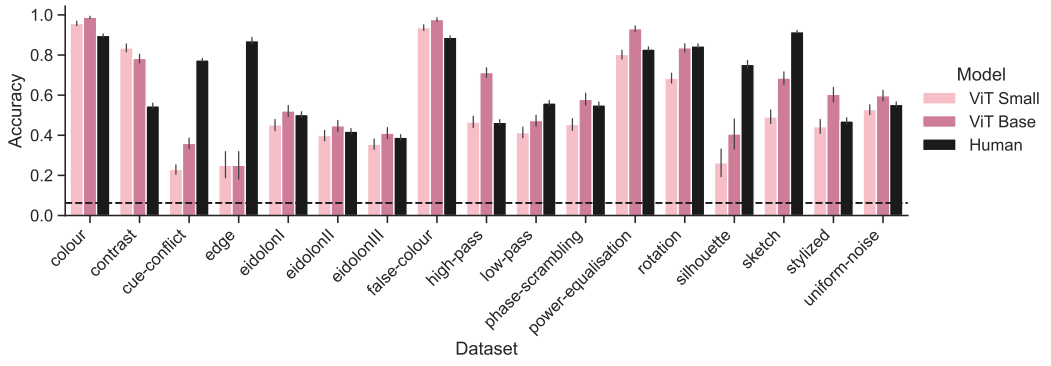
D.2 Comparison of logit lens and tuned lens

Figure 7 shows the distribution of R^2 values (predicting human DVs) achieved under logit lens and the tuned lens (Belrose et al., 2023), for the models in our experiments for which there exists a

[8] One can also interpret the direction of the $\langle -1, 1\rangle$ vector as *suppressing* the intuitive answer and the $\langle 1, -1\rangle$ as *suppressing* the correct answer. However, we do not observe these types of layers empirically.



(a)



(b)

Figure 6: Accuracy achieved by models and humans in each domain. (a) Text-based domains. For capitals recall, we show the "lenient" definition of accuracy for humans (labeled as correct by GPT-4o, which allows for minor typos). Note that we only have human data for the Competitor items in the capitals domains, while the model accuracy is shown over both the Competitor and NoCompetitor conditions. We do not have data from Llama-3.1 405B on syllogisms, nor data from humans on gender bias. (b) Vision-based tasks (Study 4).

pre-trained tuned lens. The distributions are largely similar, so we focus on results from the logit lens in the main text.

D.3 Experiment 2: Results from vision domain

Figure 8 shows the Experiment 2 results from the object recognition datasets.

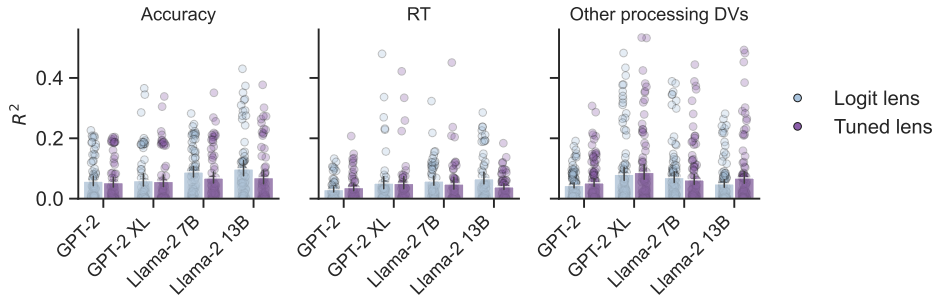
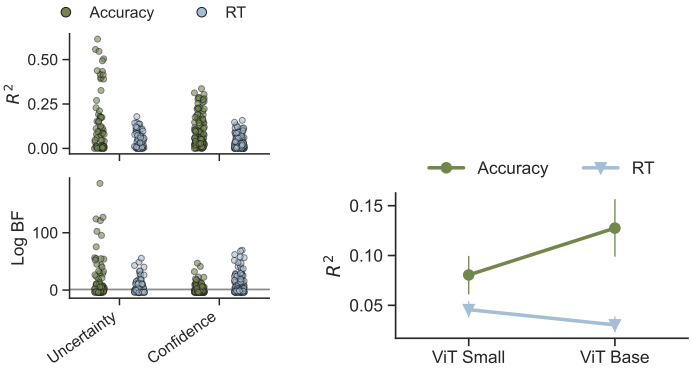


Figure 7: Distribution of R^2 values achieved under logit lens and tuned lens. Bars denote means.

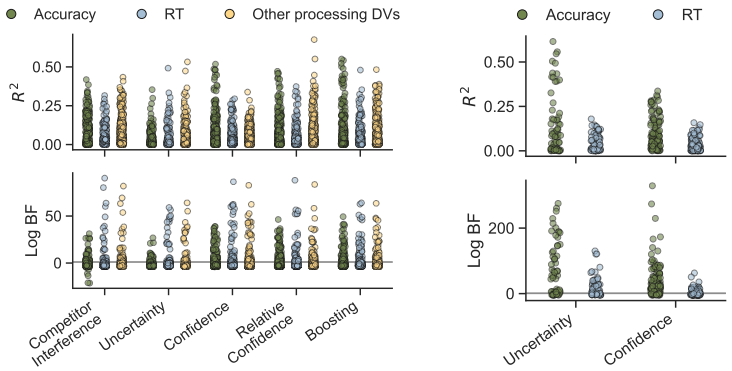


(a) (b)

Figure 8: **Experiment 2 results for vision domain.** (a) Top: R^2 achieved by model processing IVs across groups of human DVs. Bottom: Log Bayes Factor comparing critical to baseline regression models (final-layer). Horizontal line denotes $\log(3)$. (b) Mean R^2 across models.

D.4 Experiment 2: Log Bayes factors with respect to midpoint-layer baseline

Figure 9 shows the distribution of log Bayes Factors between critical and baseline regression models, with the static baseline measures taken from the midpoint layer of each model.



(a) (b)

Figure 9: **Experiment 2 results, relative to midpoint-layer baseline.** Log Bayes Factor (bottom facets) comparing critical to baseline regression models, where the baseline is formed by static readouts from the midpoint layer. Horizontal line = $\log(3)$. Note that the R^2 data (top facets) does not depend on baseline measures, and is identical to the R^2 data shown in Figure 3a and Figure 8a (with small visual differences due to randomness in the jitter), and is shown again here for visual comparison to the log BF results. (a) Text domains. (b) Vision domains.