Towards finding consensus about similarity of symbolic encodings associated with concepts between LLMs and human brain

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large Language Models (LLMs) and Multimodal Large Language Models (LLMs) have shown unbelievable improvement in performance in various Natural Language Understanding and Multimodal understanding tasks. The recent works evaluate representations, alignment, various of types of reasoning, grounding - text, video and audio inputs - over the tasks evaluating LLMs and MLLMs. The recent LLMs with "reasoning" or "thinking" phases generating reasoning traces (Chain-of-Thoughts or CoTs), enacting inference-time decision-making by "thinking" before generating a final response Feng et al. [2025] have shown new directions inspired from Kahneman [2011]. This approach leverages Reinforcement-Learning based finetuning along with rewards signals from variants of reward models while scaling up test-time compute. This paper re-introduces the previously examined individual findings from few different works Silver and Mitchell [2023] Pavlick [2023], Shani et al. [2025], Geh et al. [2024b], Opedal et al. [2024], Saparov and He [2023] and more. This paper attempts to find if there is a common consensus about similarity of symbolic encodings between LLMs and human brain. The symbolic encodings refer to alignment of symbols (words and sentences) w.r.t concepts, conceptual categories and conceptual structures.

1 Introduction

Earliest Connectionist systems have been developed from the idea to represent information in the form of the tensor product vectors to capture the representations of symbolic structures Smolensky [1987] Piantadosi et al. [2024]. The goal of Cognitive architectures is to replicate human cognition Saparov and Mitchell [2022]. However recent studies that attempt to answer "what are the Large Language Models (LLMs) supposed to model" Blank [2023] suggest lack of consensus from perspectives from within cognitive sciences studies. Recent cognitive science studies note that the LLMs operate at subsymbolic level, similar to humans, as reiterated from Silver and Mitchell [2023] that symbols are characterizations of subsymbolic processes of thought and this in itself makes symbols crucial for intelligent systems. There is a greater emphasis on LLMs capturing the human-like encodings of symbolic and conceptual information, and their relevance to reasoning.

Large Language Models (LLMs) have shown exceptional improvement in reasoning and semantic understanding, on Natural Language Understanding tasks OpenAI et al. [2024b]. LLMs also expand their language learning capabilities to translate/convert a natural language textual input into a programming languages such as C, Java and Python Brown et al. [2020]. Saparov and He [2023] and Olausson et al. [2023] involving First Order Logic aligned methods suggest LLMs fail at planning stages and that they suffer from "fallacy of the converse" in Natural Language Inference (NLI) tasks. Vision Large Language Models (VLLMs) have extended the capabilities over Multimodal reasoning

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

tasks such as Visual Question Answering (VQA), Vision Language Retrieval (VLR) and Natural Language for Visual Reasoning (NLVR) than the other Multimodal AI systems Manzoor et al. [2023]. More recent GPT-4o can also effectively process the textual, visual and audio inputs OpenAI et al. [2024a]. The recent work to evaluate if LLMs can learn low-resource languages using In-Context Learning (Zhang et al. [2024] and creating Constructed Languages by decomposing language design into phases using an LLM pipeline Alper et al. [2025] demonstrate advancement in new language learning and creation (*known as Computational conlanging*) tasks. Grounding conceptual spaces in language-only models Patel and Pavlick [2022] has also been evaluated while Multimodal Large Language Models (MLLMs) struggle with OCR-scanned documents for visual text grounding Li et al. [2025] though MLLMs have improved on multi-modal understanding tasks Zhao et al. [2023]. With LLMs' and MLLMs' increasing availability to public and subsequent rise in the usage by the public, recent works such as enabling unbiased discourse with mediation during democratic deliberation Tessler et al. [2024], evaluation of Theory-of-Mind (ToM) concepts via social reasoning capabilities in LLMs Gandhi et al. [2023], implicature-based inference in pragmatic understanding Ruis et al. [2023], human-like affective cognition in LLMs Gandhi et al. [2024] have evaluated high-level cognitive behaviors and various interesting applications of LLMs.

2 Similarity of symbolic encodings between LLMs and Humans

To discuss about the similarity of symbolic encodings that focus on concepts, between humans and LLMs, I bring together few works to attempt to verify similarity of symbolic encodings and also present a potential counter argument.

Recent works on the role of symbols Silver and Mitchell [2023] hypothesize symbols characterize sub-symbolic processes that help to communicate a thought. Silver and Mitchell [2023] distinguishes between symbolic representations and concept representations and proposes and asks some key Neuro-Symbolic questions. Silver and Mitchell [2023] attempts to question if symbols play an internal role for the agent beyond external communication such as contributing towards agent's reasoning processes in learning, memory storage and retrieval. For context, conrep (concept's neural representation) is agent's internal neural activity that encodes concept referred by a symbol and symrep (symbol's neural representation) agent's internal neural activity that encodes symbol. Silver and Mitchell [2023] suggests that concept representations are associated with a symbol - where a symbol can be an English word, Portugese word or a picture and further describes properties of LLMs by drawing analogy between symreps and conreps. Some of the properties of LLMs as noted in Silver and Mitchell [2023] suggests LLMs represent conreps of words and sentences (where words, sentences are symbols) in the form of vectors of neural activations, such as Word embeddings that capture the meaning of the words which may be used to predict the neural activation of individual words in human brain. The authors also suggest the similarity of symbol encodings from LLMs to humans and viceversa. LLMs process each word in the input (symbols) and generate an associated conrep (the neural activation) by learning which other words in the input it needs to give "attention" to, a mechanism used by an autoregressive model of word sequences. The transformer architecture explores which other words in the textual input are most relevant to modify the conrep associated with current word and then determines how to modify conrep (add's a learned vector to current word's conrep). Other properties of LLMs are that they learn to modify context-free conreps associated with individual words by taking into account the specific context of the sentence containing the word. I present some of the findings from Silver and Mitchell [2023] from comparison between human brain neural activations to that of LLMs' for symbolic and neural representations, as follows:

1. Consistent encodings of symbols upon reading the same word leads to "repeatable distributed patterns of neural activity/vectors of neural activity" Silver and Mitchell [2023].
2. Encodings of symbols focus on concepts, meaning patterns of neural activity associated with symbol stimuli (such as "cat") describe its associated concept (conrep), not just its symbol (symrep), including sound of the word cat, the images of cat, even sense of touching a cat.

2.1 Text-only LLMs and similarity of symbolic encodings

Connecting the dots from Silver and Mitchell [2023] with findings from Pavlick [2023] suggests that text-only LLMs, despite lack of groundings, are able to grasp conceptual structure of language. Grounding defined as "the ability to tie a word for which they have learned a representation to its referent in the non-linguistic world" Pavlick [2023]. The analyses and emphasis on symbols and grounding in Language Models Pavlick [2023] in text-only LLMs suggests conceptual structures are captured and how they can be leveraged for mapping LLMs to grounded conceptual spaces, even without built-in multi-modal understanding in LLMs such as in GPT-2 and GPT-3. Drawing upon this idea, the contextual information and conceptual structures learnt by the words such as color or direction, indicate that extent to which conceptual structure of LLMs reflects the conceptual structure of non-linguistic world. For example, the inputs are in textual format containing what "left" means in a textual description of gridworld. A key finding of Pavlick [2023] is that LLMs tend do well on these example tasks even in isomorphic rotated worlds and they are not using naive memorisation to succeed in such tasks. Further authors suggest that such learnt conceptual structure can be used to ground by leveraging data-efficient approaches. The authors note the limitation of text-only LLMs that for complex visual inputs grounding, a textual input depiction is unlikely to have required grasp of complex conceptual structures inherent within such visual inputs (non-linguistic world).

Combining one of the key findings from Silver and Mitchell [2023], the question is: do the findings from Pavlick [2023] that linguistic world models tend to capture conceptual structures, might suggest encodings focus on concepts to describe symbol stimuli such as "left" describe its associated concepts in the textual gridworld and its related concepts of "right" in the same gridworld?

2.2 Using compression-meaning tradeoff evaluation as a measure for comparing similarity of symbolic encodings

Another interesting work Shani et al. [2025] discussed on how humans and LLMs trade-off between compression and preservation of semantic meaning and explores if LLMs develop conceptual structures analogous to human cognition. The authors further suggest human concepts balance semantic richness and cognitive manageability suggestive of Information compression. This resonates with the ideas from Silver and Mitchell [2023] that suggest symbols are characterization of our thoughts that allows to explain subsymbolic thinking to ourselves and others (synonymous to "reasoning traces", "thinking" phases in LLMs) and act as constraints for inference and learning about the world. There are three key research questions in Shani et al. [2025] that ask 1) "how do emergent concepts in LLMs align with human-defined conceptual categories?" Shani et al. [2025] referred to as *representational compactness*, 2) "do LLMs and humans exhibit similar internal geometric structures within these concepts?" Shani et al. [2025] *semantic preservation* and 3) "how do humans and LLMs differ in their strategies for balancing meaning preservation and information compression?" Shani et al. [2025] referred as *"compression-meaning tradeoff" measure*. The authors further draw on the Information theoretical constructs such as Rate-Distortion measure Theory (RDT) and Information Bottleneck principle (IB), where rate R is the representational complexity needed to represent source X as C where R is subjected to maximum distortion D (fidelity loss, w.r.t semantic preservation). The goal is to optimize $R + \lambda D$ for evaluation of representational efficiency. More details about how IB, RDT and objective are formulated in A.2. The findings pertaining to semantic preservation suggest there is above-chance alignment with human conceptual categories and do not fully mirror nuanced prototype structures evident in human typicality judgments. The typicality here refers to "robin" as a "typical bird", "bat" as atypical (because it is a mammal within the context of comparative human conceptual category of "bird") Shani et al. [2025]. Lastly the compression-meaning tradeoff evaluation results suggests superior information-theoretic efficiency in LLMs' conceptual representations in comparison to human conceptual structures, which strongly suggests divergence in the strategies used for balancing informational compression with semantic meaning. This suggests that there is similarity in alignments with human conceptual categories and ability to recover human-like categories from their item embeddings, though both employ different strategies for compression-semantic balance. The idea focuses using information-theoretic measures for measuring and comparing compression-semantic tradeoffs between both humans and LLMs for alignment with human conceptual categories, which is a different method to that of methods used in Silver and Mitchell [2023] and Pavlick [2023].

Based on the findings from Silver and Mitchell [2023] there seems to be agreement with that of the findings in Shani et al. [2025] that there is similarity and presence of alignment between human conceptual categories with that of items/tokens as symbols/symbol stimuli (by information-theoretic measures), thought there is divergence between the strategies employed by LLMs and humans for compression-meaning tradeoffs exists. There seems to be potential consensus that LLMs do tend to capture conceptual structures relevant to symbols/tokens for current token embeddings, similar to that of humans.

2.3 Works regarding hallucinations, cognitive biases, logical reasoning and more

There are several works examining the presence of cognitive biases Gupta et al. [2023] Opedal et al. [2024], logical fallacies Lalwani et al. [2024] addressing hallucinations, Li et al. [2024] Gu et al. [2025] reviews LLMs as judges, RAG-based methods Fan et al. [2024], Feng et al. [2024] that suggest LLMs replicate inherent biases, beliefs of the humans. Recent works on Wang et al. [2024] suggest various aspects of human-centric perspectives and analyze cognition at individual and collective spaces respectively. The evaluation of deductive reasoning through First order logic based tasks to derive and evaluate conclusions from logical premises Saparov and He [2023] which suggests room for improvement in planning for LLMs. These works also have potential to resonate similarity of symbolic encodings, since humans themselves suffer from these potential drawbacks, LLMs resonate these common drawbacks from the humans Krawczyk [2017].

3 Alternative Views

3.1 Insights from Non-Canonical tokenization for similarity of symbolic encodings

Another recent work Geh et al. [2024a] examines tokenization encodes text into canonical tokens. The authors discuss importance of Byte-Pair Encodings (BPE) commonly used in LLMs which repeatedly merges most frequent byte pair of tokens into the new token into merge table and each one of the entries are assigned priorities. The results are further processed by splitting the string into constituent characters, then combines the pair of tokens that have highest priority merge rule which results in canonical tokenization. The BPE method with dropout adds more mass to non-canonical tokenizations. The key finding from the authors are that majority of non-canonical tokenizations belong to non-English cases with large portions consisting of code and other language tokens which is word-dependent. Some of the non-canonical tokenizations contain grammatically correct words, but longer texts seem to contain some mass attributed to non-canonical tokenizations. The authors extend this interesting finding to verify that instead of using single tokenization, if aggregating over all tokenizations each weighted by their probability, then to compute marginal probability of the string x. Their study showed there is most of the probability mass over canonical tokenization. These findings and further experimentation showed improvement in accuracy for non-canonical tokenization concluding there is signal in the *non-canonical tokenization* space suggesting they retain meaningful information.

At the language level, the findings from Geh et al. [2024a] show that non-canonical tokenizations capture more meaningful information, under controlled configurations for downstream tasks such as Question and Answer datasets (Q&A). How can we map non-canonical tokens that seem to capture meaningful information in Q&A tasks with that of symreps and conreps from human brain imaging in an analogous Q&A task ? Does this suggest that the canonical vs non-canonical tokenization techniques can lead to vectors of activations until "symbol stimuli" in LLMs? Example: $[B, a, t]$ are canonical tokens seen by an LLM during tokenization until LLM searches and finds relevant tokenizations and it's associated concept of "Bat", within the context (whether it's a baseball bat or "Bat" as a "mammal"), while humans receive symbol stimuli for *Bat* as an English word, or as a picture based on symreps and conreps respectively. Do these present similarities of encodings of symbols at conceptual level or do we need to find another meeting point to compare the similarities of symbolic encodings ?

3.2 Conclusion

This work broadly studies key works for consensus towards similarity of symbolic encodings between LLMs and human brain, by drawing inspiration from Silver and Mitchell [2023]. Combining the aforementioned analyses, there are sufficient evidences that suggest similarity exists. Two contrasting research questions emerge that - to what extent is the similarity acceptable VS to what extent is the dissimilarity desirable VS domain/task-based expert human cognition (finetuned) VS general human cognition (pretrained) - seems to be restricting where to draw the line for consensus over similarity of symbolic encodings between LLMs and human brain.

References

Morris Alper, Moran Yanuka, Raja Giryes, and Gašper Beguš. Conlangcrafter: Constructing languages with a multi-hop llm pipeline. *arXiv preprint arXiv:2508.06094*, 2025.

Idan A Blank. What are large language models supposed to model? *Trends in Cognitive Sciences*, 27 (11):987–989, 2023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL `https://arxiv.org/abs/2005.14165`.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501, 2024.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*, 2024.

Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. Efficient reasoning models: A survey, 2025. URL `https://arxiv.org/abs/2504.10903`.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529, 2023.

Kanishk Gandhi, Zoe Lynch, Jan-Philipp Fränken, Kayla Patterson, Sharon Wambu, Tobias Gerstenberg, Desmond C. Ong, and Noah D. Goodman. Human-like affective cognition in foundation models, 2024. URL `https://arxiv.org/abs/2409.11733`.

Renato Lui Geh, Honghua Zhang, Kareem Ahmed, Benjie Wang, and Guy Van den Broeck. Where is the signal in tokenization space? *arXiv preprint arXiv:2408.08541*, 2024a.

Renato Lui Geh, Honghua Zhang, Kareem Ahmed, Benjie Wang, and Guy Van den Broeck. Where is the signal in tokenization space?, 2024b. URL `https://arxiv.org/abs/2408.08541`.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL `https://arxiv.org/abs/2411.15594`.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*, 2023.

Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

Daniel Krawczyk. *Reasoning: The neuroscience of how we think*. Academic Press, 2017.

Abhinav Lalwani, Lovish Chopra, Christopher Hahn, Caroline Trippel, Zhijing Jin, and Mrinmaya Sachan. Nl2fol: translating natural language to first-order logic for logical fallacy detection. *arXiv preprint arXiv:2405.02318*, 2024.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024. URL `https://arxiv.org/abs/2412.05579`.

Ming Li, Ruiyi Zhang, Jian Chen, Jiuxiang Gu, Yufan Zhou, Franck Dernoncourt, Wanrong Zhu, Tianyi Zhou, and Tong Sun. Towards visual text grounding of multimodal large language model, 2025. URL `https://arxiv.org/abs/2504.04974`.

Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20 (3):1–34, 2023.

Theo X Olausson, Alex Gu, Benjamin Lipkin, Cedegao E Zhang, Armando Solar-Lezama, Joshua B Tenenbaum, and Roger Levy. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. *arXiv preprint arXiv:2310.15164*, 2023.

Andreas Opedal, Alessandro Stolfo, Haruki Shirakami, Ying Jiao, Ryan Cotterell, Bernhard Schölkopf, Abulhair Saparov, and Mrinmaya Sachan. Do language models exhibit the same cognitive biases in problem solving as human learners?, 2024. URL `https://arxiv.org/abs/2401.18070`.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho

Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024a. URL https://arxiv.org/abs/2410.21276.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon

Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024b. URL `https://arxiv.org/abs/2303.08774`.

Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. In *International conference on learning representations*, 2022.

Ellie Pavlick. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(2251):20220041, 2023.

Steven T Piantadosi, Dyana CY Muller, Joshua S Rule, Karthikeya Kaushik, Mark Gorenstein, Elena R Leib, and Emily Sanford. Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, 28(9):844–856, 2024.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms, 2023. URL `https://arxiv.org/abs/2210.14986`.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought, 2023. URL `https://arxiv.org/abs/2210.01240`.

Abulhair Saparov and Tom M. Mitchell. Towards General Natural Language Understanding with Probabilistic Worldbuilding. *Transactions of the Association for Computational Linguistics*, 10: 325–342, 04 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00463. URL `https://doi.org/10.1162/tacl_a_00463`.

Chen Shani, Dan Jurafsky, Yann LeCun, and Ravid Shwartz-Ziv. From tokens to thoughts: How llms and humans trade compression for meaning, 2025. URL `https://arxiv.org/abs/2505.17117`.

Daniel L Silver and Tom M Mitchell. The roles of symbols in neural-based ai: They are not what you think! In *Compendium of Neurosymbolic Artificial Intelligence*, pages 1–28. IOS Press, 2023.

Paul Smolensky. Analysis of distributed representation of constituent structure in connectionist systems. In D. Anderson, editor, *Neural Information Processing Systems*, volume 0. American Institute of Physics, 1987. URL `https://proceedings.neurips.cc/paper_files/paper/1987/file/66f041e16a60928b05a7e228a89c3799-Paper.pdf`.

Michael Henry Tessler, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes, et al. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719): eadq2852, 2024.

Jing Yi Wang, Nicholas Sukiennik, Tong Li, Weikang Su, Qianyue Hao, Jingbo Xu, Zihan Huang, Fengli Xu, and Yong Li. A survey on human-centric llms, 2024. URL `https://arxiv.org/abs/2411.14491`.

Chen Zhang, Mingxu Tao, Quzhe Huang, Zhibin Chen, and Yansong Feng. Can llms learn a new language on the fly? a case study on zhuang. In *The Second Tiny Papers Track at ICLR 2024*, 2024.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms, 2023. URL `https://arxiv.org/abs/2307.08581`.

A Technical Appendices and Supplementary Material

A.1 About properties of LLMs discussed in Silver and Mitchell [2023]

The two additional properties not included in aforementioned discussion, due to focus on "language" relevant similarity comparisons:

1. Encodings are multi-modal, meaning representations in human brain and Artificial Neural Networks "get similar patterns" whether hearing or writing, word could be in English or Portugese, "where full representations are spread across sensory and motor modalities" Silver and Mitchell [2023].
2. Dual Architecture draws upon Kahneman's theory of thinking fast and slow, by using two systems named: System 1 that thinks fasts and System 2 that thinks slow by applying rules, logic and evidences. "Kahneman's theory suggests brain learns quickly to activate a neural pattern Y, if it was frequently coactivated with neural pattern X" Silver and Mitchell [2023]. For example even if the image of "peach" or symrep of "peach is partial or vague such as canned peaches, or smashed peaches, brain generates conrep of typical peach. This seems to be modulated by grounded perception.

A.2 Formulations for compression-meaning tradeoff evaluation as a measure for comparing similarity of symbolic encodings between LLMs and humans Shani et al. [2025]

IB seeks a compressed representation C of an input X that maximizes information about relevant variable Y minimizing $I(X;C)$, mutual information C retains about X, is the bottleneck cost. The goal $\mathcal{L}$ to balance RDT's rate and distortion, $\mathcal{L}$ designed to explicitly balance complexity term R, representing X through conceptual clusters C. By RDT theory, $X : X, x_1, x_2 \cdots \in X$ is a source sequence. The reproduction sequence is a potential output $\hat{X} : \hat{X}, \hat{x_1}\hat{x_2} \cdots \in \hat{X}$ and the distortion measures the loss or distance (that are normalized/normal distortion measures). The distortion measures, for RDT for the the goal $\mathcal{L}$, is for semantic information lost or obscured within the clusters (variance of each $x_i \in X$ embeddings relative to concept cluster centroids). To combine the three research questions with RDT and IB formulations, where X are the token embeddings, $\mathcal{L}(X, C; \beta) = \textit{Complexity}(X, C) + \beta \cdot \textit{Distortion}(X, C)$ The further formulations $I(X; C) = H(X) - H(X \mid C)$ applies to initial and conditional entropies respectively. If Cluster assignments C make the specific items X more predictable, then that signifies greater compression. The complexity *Complexity*(X, C) and its details of formulations expressed in terms of respective entropy formalizes representational compactness. The Distortion term *Distortion*(X, C) measures loss of sematic fidelity incurred by grouping items into clusters, which is measured as average intra-cluster variance of the item embeddings and this formalizes semantic preservation. The unified objective $\mathcal{L}(X, C; \beta)$ combines *Complexity*(X, C) i.e. representational compactness and *Distortion*(X, C) i.e. semantic preservation together formalizes compression-meaning tradeoff.

Further by using k-means and other relevant applicable metrics, the findings relevant to representational compactness suggest above-chance alignment with human conceptual categories and can recover human-like categories from their embeddings.

A.3 Details about Canonical vs non-canonical tokenization problem formulation

For each string x, canonical tokenizer yields a canonical tokenization vector v^*, which is evaluated by LLM for canonical probability $p(v^*, x)$, which is one of the exponential number of possible tokenizations and space. The distribution of tokenization space from Geh et al. [2024a] shows larger probability mass on the canonical tokenization.

For computing marginals, the findings suggest that for short strings, the results seem to show that canonical probability is close to true marginal. For the longer texts, the approximate marginal approaches closer to canonical tokenization probability.