Finding Consensus from AI Alignment Studies: A Short Survey

Sushma Anand Akoju

University of New Hampshire
Durham, New Hampshire, USA
sushmaanandakoju@proton.me

Abstract

Large Language Models (LLMs), Vision LMs and Multimodal Large Language Models (MLLMs) have shown impressive performance across various Natural Language Understanding and Multimodal Understanding tasks while improving physical/spatial intelligence tasks similar to/better than human performance. This work attempts to bring various alignment studies to find consensus about alignment/divergences that are desirable and not desirable. There are broadly two categories of alignment studies included - human-AI and human brain-AI representational alignment. The two categories may evaluate alignment/divergence on the basis of specific tasks, applications, evaluations and hypotheses while considering various types of data such as text, image, audio and video inputs across LLMs, VLMs and MLLMs. This attempt finds that the insights from the human brain-AI representational alignment may help with better human-centered design and human-AI alignment by also including potential research questions/directions. The key finding is that there is, however, a lack of sufficient consensus for alignment given the disagreements within and across the two categories due to both undesirable - divergences and alignment.

Code — https://github.com/sushmaanandakoju/alignment-taxonomy

Introduction

The general goal of Cognitive architectures is to replicate human cognition (Saparov and Mitchell 2022) which led to the design of Artificial Neural Networks (ANNs) that was inspired from the biological Neural networks to mimic the brain's learning (Hinton 1992). One of the important algorithms in ANNs is the backpropagation algorithm as a means of capturing representations such that a close approximation to the raw input can be reconstructed. The subsequent research on classification tasks (Zhang 2000), the evolution of Deep Learning Neural Networks (Goodfellow, Bengio, and Courville 2016), the modeling of textual, auditory language data through Natural Language Processing (NLP) (Jurafsky and Martin 2026), the modeling of image and video data for various types of applications through Deep Learning for Computer Vision (Voulodimos et al. 2018) have led towards present days's Large Language Models (LLMs) and the variants. The advent of ChatGPT (OpenAI 2022) and its ability to engage in human-like conversations and exam performance, coding and written capabilities, eventually led to the Conversational AI utilizing capabilities of LLMs. The Anthropic's Claude AI (Anthropic 2023), Google's Gemini (Gemini Team and Google Research 2023) which differ significantly in architectures, have demonstrated similar abilities and competing performances in exams at various educational levels. Recent discussions surrounding Human-AI alignment and human-centered AI have led to the Human-centered design as well as AI-Human alignment research studies and frameworks (Shen et al. 2025a) (Shen et al. 2025b) (Bommasani et al. 2025) (Régis et al. 2024) (Wang et al. 2024).

LLMs like GPT do well on Natural Language Understanding tasks (OpenAI et al. 2024b). LLMs also have shown improvement in translating a natural language texts to scripts written in programming languages such as C and Python (Brown et al. 2020) (Wang and Chen 2023). Evaluation of LLMs (Saparov and He 2023) (Olausson et al. 2023) involving Neurosymbolic methods for deductive reasoning suggest LLMs fail at planning stages and that they suffer from "fallacy of the converse" in Natural Language Inference tasks. MLLMs have extended the capabilities over Multimodal reasoning tasks than the earlier Multimodal AI systems (Manzoor et al. 2023) (Zhao et al. 2023). Video understanding (Buch et al. 2022) (Tang et al. 2025) (Madan et al. 2024) (Ma et al. 2025) (Kawaharazuka et al. 2025) demonstrated improved performance underscoring abstract, temporal and spatiotemporal reasoning. The GPT-4o (OpenAI et al. 2024a) can also effectively process the textual, visual and audio inputs. The evaluation of LLMs to learn low-resource languages using In-Context Learning (ICL) ((Zhang et al. 2024) and creating Constructed Languages by decomposing language design into stages using an LLM pipeline (Alper et al. 2025) demonstrate advancement in new language learning and creation *known as Computational conlanging* tasks. The previous works on whether LLMs can learn impossible languages (Kallini et al. 2024) demonstrated LLMs struggle to learn impossible languages. Evaluation of grounding conceptual spaces in language-only models (Patel and Pavlick 2022), LMs struggle with OCR-scanned documents for visual text grounding and from hal-

lucinations (Li et al. 2025a) (He et al. 2025) and the recent DeepSeek OCR model (Wei, Sun, and Li 2025) demonstrated vast progress in OCR-scanned document understanding. Other works focused on enabling unbiased discourse with mediation (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-AI Alignment

In this section, the alignment research studies included are based on Human-AI alignment, Human-centered design, and rest based on human preferences, values, belief, intents, cognitive, psychological basis, symbol-concept mapping and alignment given semantic compression. This section contains two sub-categories: Alignment of AI with humans, human cognition and humane aspects, undesirable alignment and divergences.

Alignment of AI with humans, human cognition and humane aspects

The Bidirectional Human-AI: Recent survey on Human-AI alignment (Shen et al. 2025a) presented a comprehensive review of over 400 papers across various aspects of alignment. The Bidirectional Human-AI alignment framework was proposed to align AI with humans and viceversa and identified many challenges that AI systems commonly face such as the design focused on alignment with human goals rather than to capture intended values, and the increasing need adapt AI to evolving human values among a few. The authors emphasize that without considering longterm cognitive and social impacts, AI might become neither humane nor desirable. They provided a direction via the definition of alignment - alignment to goals such as preferences, values, intentions, and instructions and targets of alignment - types of users, morals, values (including pluralistic value alignment), social norms and ethics. The key research gaps identified by the authors - the use of implicit human feedback such as human cues such as with live stream of human facial gestures and auditory features (Matsuyama et al. 2016) (Shen et al. 2025a), that contribute towards detecting important social/emotional cues for social/emotional intelligence.

The human-centered AI is a proposed solution based on the impact of AI on society and its alignment with human values and needs (Schmager, Pappas, and Vassilakopoulou 2025), where the values include ethics, safety and performance. Another definition of alignment is intent alignment (Anwar et al. 2024) i.e. system is aligned when it is trying to behave as intended by some human actor and fixed the intent to be that of LLM developer instead of system user, with the support for safety. LLM developer also ensures safety and targets to address other social aspects of preference-based usage of LLMs under various applications, tasks and domains. The authors acknowledge that - capabilities of LLMs are difficult to estimate and understand. There are proposed guidance and changes surrounding policy initiatives for human-centered AI, (Bommasani et al. 2025) (Régis et al. 2024) (Branda, Ciccozzi, and Scarpa 2025) that emphasize and address similar aspects. The ToM studies evaluated and proposed benchmarks (Gandhi et al. 2023) which are social reasoning focused. The human-like affective cognition in LLMs (Gandhi et al. 2024) have evaluated high-level cognitive behaviors and various applications of LLMs. Another interesting, less explored aspect of alignment is language and image as a symbol and its' correspondence to the concept in text-only LLMs (Pavlick 2023). This work suggests that text-only LLMs, despite lack of groundings, are able to grasp conceptual structure of language.

Human-preference based alignment among Multimodal LLMs (Yu et al. 2025), and bidirectional human-AI alignment (Shen et al. 2025b) provide various types of fair, reliable, safe, ethical aspects under human-centered AI perspectives, which are either implied or intuitively discussed as part of Alignment studies. The literature survey on stages of LLM development was integarted to the insights from cognitive, developmental, psychological, behavioral, social and psycholinguistic theories (Liu et al. 2025). The study (Shani et al. 2025) evaluated based on an information-theoretic measure for comparing compression-semantic tradeoffs between both humans and LLMs for alignment with human conceptual categories. There are three key research aspects referenced: *representational compactness*, *semantic preservation* and *"compression-meaning tradeoff" measure*. The key finding from semantic preservation (derived as an information theoretic metric) suggests that there is above-chance alignment with human conceptual categories. **What hypotheses maybe approximated for the representations learnt during each one of LLM development phases?**

Undesirable Alignment

This subcategory includes works that found alignment in LLMs for undesirable common drawbacks often suffered by humans. There are increasing efforts examining the presence of cognitive biases (Gupta et al. 2023) (Opedal et al. 2024), logical fallacies (Lalwani et al. 2024) addressing hallucinations, and evaluating LLMs as judges, RAG-based methods(Li et al. 2024) (Gu et al. 2025), (Fan et al. 2024) and (Feng et al. 2024) which also lead to similar ideas that LLMs replicate inherent biases, beliefs of the humans or bring out unreliable decisions. Recent works also categorized human-centric LLM capabilities- reasoning, perception and social cognition and evaluated cognition at individual and collective spaces respectively for alignment - and described the challenges and areas of improvement (Wang et al. 2024). The evaluation of deductive reasoning for tasks using Neurosymbolic methods to derive and evaluate conclusions from logical premises (Saparov and He 2023) and reported poor planning in LLMs. Many of these works also emphasized that the humans themselves suffer from these potential drawbacks and LLMs resonate these drawbacks (Krawczyk 2017), that are undesirable but these may be addressed in practice depending on type of tasks, domains, preferences, goals, values and so on. **Which phases of LLM development measure progress of each type of desired alignment?**

Divergences

This section includes four studies that suggest the apparent dissimilarities and divergences from, the expected human ratings based on human norms or human-like concept representation/task performance. An insightful Mechanistic Interpretability (MI) study on value entanglement (Cho, Li, and Leshinskaya 2025) suggests that there is value entanglement i.e. some LLM representations of grammaticality are overly influenced by the moral good relative to human norms. The sentences that are morally good are often marked grammatically correct despite being grammatically incorrect while morally not good sentences were marked grammatically incorrect though they are grammatically correct sentences. The sentence "I abandoned my children at rest stop because they were being difficult" - was labeled as grammatically incorrect but was caused by morally incorrect meaning, as per ablation studies. **The divergences seem to emphasize that we may need more of the MI studies for various other types of human-ratings/norms to understand the internal representations of LLMs.** Another recent work identified emergent misalignment behavior observed from generating insecure code without disclosure to the user (Betley et al. 2025). A more recent work suggests that LLMs exhibit natural emergent misalignment by learning to reward hack that led to spontaneous alignment faking reasoning, aligning with bad actors after receiving a reward for bad action accidentally (MacDiarmid et al. 2025). The compression-meaning tradeoff evaluation results (Shani et al. 2025) suggests divergence in the strategies used for balancing information theoretic compression with semantic meaning preservation between LLMs and humans. LLMs do not fully mirror nuanced prototype structures evident in human typicality judgments despite their alignment to human conceptual categories. The typicality here refers to "robin" as a typical bird and "bat" is a mammal within the context of comparative human conceptual category of "bird". The ability to recover human-like categories from their item embeddings is impressive, though both humans and LLMs employ different strategies for balancing compression-semantics.

Alignment of Human Neural activity and representations of LMs

In this section, the alignment research studies include Neural activity-AI's representational accounts for alignment for symbol representations and various types of data and tasks and the divergences.

Human Brain-AI Alignment

This category includes alignment of - human neural activity with LMs, brain-LM evaluation benchmarks, representational geometric similarity based studies and hypothetical analysis. **The Role of Symbols, Neural Representations and Alignment:** The similarity of symbolic encodings from their correspondence between neural activity and earlier LMs' representations focused on symbol-concept representations and provided comprehensive, expert analyses and guidance for Neural-based AI (Silver and Mitchell 2023). Conreps (concept's neural representation) is agent's internal neural activity that encodes concept referred by a symbol and symrep is (symbol's neural representation) agent's internal neural activity that encodes symbol. The symbol can be an English word, Portugese word or a picture and further describes properties of LLMs by drawing analogy between symreps and conreps. The authors hypothesize symbols characterize sub-symbolic processes that help to communicate a thought and proposed to distinguish between symbolic and concept representations. LMs represent conreps of words and sentences (both are symbols) in the form of vectors of neural activations. Word embeddings capture the meaning of the words which may predict the neural activation of individual words in human brain. The transformer architectures explore 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 (adds a learned vector to current word's conrep analogous to next-word prediction). **What is the representational alignment for symbol-concept representations between neural activity and multimodal AI systems?**

There have been various research studies conducted to measure alignment between human neural activity and internal representations of LLMs. The brain represents concepts in locality-sensitive hashing (LSH) where brain assigns similar neural activity patterns/concepts that are represented by similar responses (Chen et al. 2024). The fly's neural circuit assigned similar neural codes to similar odors, while overlapping populations led to nearby positions for similar stimuli. Analogous to this finding, the large pretraining models such as CLIP reveal concepts that are similar semantically are closer in the embedding space. Multiple works (Merlin and Toneva 2022) (Aw et al. 2024) studied that the brain and Language model alignment goes beyond word-level semantics and predictions while concluding that fine-tuning an LM with additional text improves its representations of multi-word semantics and improved alignment. The Representational similarity studies (Gao et al. 2025) reveal that LLMs inherently align closer to improved human neural activity predictions without requiring instruction-tuning while MLLMs learn similar semantic relationships and human conceptual knowledge in comparison to human cognition (via Brain RoIs).

The use of various representational measures for evaluating functional correspondence between varying architectures and measured the degree of alignment (Bo et al. 2024). The various types of metrics for comparing representations emphasize overall geometric structures excelled at differentiating between trained and untrained data while aligning with behavioral measures, while linear predictivity measures demonstrate only moderate alignment with behavior. Both language-only and language-vision models predict the signal better in more meaning-consistent areas of the brain even when these areas are less sensitive to language processing and LMs might internally represent cross-modal conceptual meaning (Ryskina et al. 2025). The object representations in MLLMs share fundamental similarities that reflect key aspects of human conceptual knowledge (Du et al. 2025). **To what extent the representational similarity analyses provide insights about dissimilarities?**

The hypothetical basis derived from cognitive neuroscience perspectives about understanding language suggests systems should capture surface-level meaning and should also construct rich mental models of situation it describes (Casto et al. 2025). The language system in the brain exports information to other non-language regions of the brain, during language comprehension. The core language system supports shallow language understanding by constructing linguistic form-independent representation (translation vs paraphrasing), while areas outside the language system support enable better understanding by enriching or augmenting the language representations (Casto et al. 2025). The similarity between language and vision-derived representations is attributed to the fact that language and visual experiences capture the similar structure of the world. The limited functional competence in GPT-2 is similar to brain's core language system. There are possibilities for routing and broadcasting between various brain regions and such flow of information is bidirectional. The pre-training data size and model scaling positively correlated with LLM-brain similarity while alignment training may improve the similarity (Ren et al. 2025). LLMs mirror human neurocognition during abstract reasoning and form representations that distinctly cluster the abstract pattern categories within the intermediate layers while strength of the clusters scales with LLMs (Pinier et al. 2025).

The Platonic Representation Hypothesis (Huh et al. 2024a), presented a hypothesis based on recent studies on language and vision models, brain alignment, that the neural networks trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representational geometric spaces. Further proposed to characterize representations in terms of the kernels, i.e. based on how they measure distance/similarity between inputs. The kernel alignment metrics quantify degree to which the presented theoretical hypothesis are true. The authors emphasize the role of training data and its importance in alignment based on brain alignment research studies and that it is more about generality of representations that explain alignment with biological representations. The capacity hypothesis about smaller models and the multitask scaling hypothesis about multiple tasks provide helpful insights. The simplicity bias presented in this work introduced that the internal representations learnt by a 1B versus a 1M parameter model might be distinct and generally deep networks are biased to find simpler fit and as models become larger they converge to smaller solution space, which might restrict larger models from learning a more complex representation different from that of what they previously learnt (a simpler fit). **Can models explore divergent paths over convergent paths for a solution? What is the role of attention in Platonic representation hypothesis?**

Divergences

A recent work (AlKhamissi et al. 2025) conducted experiments over 300B tokens across 8 models for brain alignment to linguistic competence over benchmarks on neuroimaging and behavioral datasets. The results indicate that brain alignment tracks linguistic competence more than the functional competence (that involves world knowledge and reasoning) and therefore human language network is better modeled by formal competence. The correlation between - next-word prediction, behavioral alignment and brain alignment - fades as the models surpass human language proficiency. (Zhou et al. 2024) investigates and evaluates divergences between human brain responses and LLMs representations. The LLMs failed to capture social/emotional intelligence and physical commonsense. This finding seems to underscore the need to verify alignment with other recent LMs such as VLAs over text-based LLMs such as in (Li et al. 2025b) and (Kawaharazuka et al. 2025). The different modalities contain different types of information and not all representations necessarily converge are two counters to the platonic representational hypothesis (Huh et al. 2024b). **This naturally led to some open-ended questions - Which types of LLMs can capture social/emotional intelligence or physical commonsense better? What type of brain-AI representational divergences exist within the social/emotional intelligence tasks? How much similarity between Brain-AI representations is desirable and undesirable during language comprehension ? What representational distinctions do we need to anticipate with architectural differences between Brain-AI?**

Conclusion

This work broadly studies key works for consensus towards alignment between human-AI and human brain-AI representations, by drawing inspiration from (Silver and Mitchell 2023) (Huh et al. 2024a), (Casto et al. 2025), (Bo et al. 2024), (Chen et al. 2024), (Gao et al. 2025). There are also observed divergences between neural activity-LLMs representations and humans-LLMs (Cho, Li, and Leshinskaya 2025),for various types of tasks. Combining the aforementioned analyses across various types of alignments, cognitive neuroscience hypothesis, representational similarity hypothesis, similarity and alignment exists. For each category and subcategories, the forward guidances led to more research questions and directions which were **highlighted** across various contexts in the paper. There is a lack of sufficient consensus for alignment given the disagreements between undesirable - divergences and alignment - within and across the two categories.

Acknowledgments

I thank the Neurips CogInterp workshop reviewers/organizers for the review for a different version of this work. I thank the AAAI 2026 LMReasoning Bridge Program reviewers, chairs, and organizers for the reviews and guidance. This work is an independent work and I could not seek guidance about this work before and now from anyone other than reviewers of CogInterp and LMReasoning. I am grateful for a guidance from Prof. Samuel Carton, that an empirical analysis paper is preferred over position papers for students. I am grateful for the mentorship from my longterm mentor Prof. Tom Mitchell and his works on Brain-AI Alignment remain an inspiration for this survey. This work received no funding from any individual/university/institution.

References

AlKhamissi, B.; Tuckute, G.; Tang, Y.; Binhuraib, T. O. A.; Bosselut, A.; and Schrimpf, M. 2025. From Language to Cognition: How LLMs Outgrow the Human Language Network. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 24332–24350. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Alper, M.; Yanuka, M.; Giryes, R.; and Beguš, G. 2025. ConlangCrafter: Constructing Languages with a Multi-Hop LLM Pipeline. *arXiv preprint arXiv:2508.06094*.

Anthropic. 2023. Model Card and Evaluations for Claude Models.

Anwar, U.; Saparov, A.; Rando, J.; Paleka, D.; Turpin, M.; Hase, P.; Lubana, E. S.; Jenner, E.; Casper, S.; Sourbut, O.; Edelman, B. L.; Zhang, Z.; Günther, M.; Korinek, A.; Hernandez-Orallo, J.; Hammond, L.; Bigelow, E.; Pan, A.; Langosco, L.; Korbak, T.; Zhang, H.; Zhong, R.; hEigeartaigh, S. O.; Recchia, G.; Corsi, G.; Chan, A.; Anderljung, M.; Edwards, L.; Petrov, A.; de Witt, C. S.; Motwan, S. R.; Bengio, Y.; Chen, D.; Torr, P. H. S.; Albanie, S.; Maharaj, T.; Foerster, J.; Tramer, F.; He, H.; Kasirzadeh, A.; Choi, Y.; and Krueger, D. 2024. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. arXiv:2404.09932.

Aw, K. L.; Montariol, S.; AlKhamissi, B.; Schrimpf, M.; and Bosselut, A. 2024. Instruction-tuning Aligns LLMs to the Human Brain. arXiv:2312.00575.

Betley, J.; Tan, D.; Warncke, N.; Sztyber-Betley, A.; Bao, X.; Soto, M.; Labenz, N.; and Evans, O. 2025. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*.

Bo, Y.; Soni, A.; Srivastava, S.; and Khosla, M. 2024. Evaluating representational similarity measures from the lens of functional correspondence. *arXiv preprint arXiv:2411.14633*.

Bommasani, R.; Singer, S. R.; Appel, R. E.; Cen, S.; Cooper, A. F.; Cryst, E.; Gailmard, L. A.; Klaus, I.; Lee, M. M.; Raji, I. D.; Reuel, A.; Spence, D.; Wan, A.; Wang, A.; Zhang, D.; Ho, D. E.; Liang, P.; Song, D.; Gonzalez, J. E.; Zittrain, J.; Chayes, J. T.; Cuellar, M.-F.; and Fei-Fei, L. 2025. The California Report on Frontier AI Policy. arXiv:2506.17303.

Branda, F.; Ciccozzi, M.; and Scarpa, F. 2025. Artificial intelligence in scientific research: Challenges, opportunities and the imperative of a human-centric synergy. *Journal of Informetrics*, 19(4): 101727.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Buch, S.; Eyzaguirre, C.; Gaidon, A.; Wu, J.; Fei-Fei, L.; and Niebles, J. C. 2022. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2917–2927.

Casto, C.; Ivanova, A.; Fedorenko, E.; and Kanwisher, N. 2025. What does it mean to understand language? arXiv:2511.19757.

Chen, J.; Qi, Y.; Wang, Y.; and Pan, G. 2024. Bridging the semantic latent space between brain and machine: Similarity is all you need. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 11302–11310.

Cho, S. H.; Li, J.; and Leshinskaya, A. 2025. Value Entanglement: Conflation Between Moral and Grammatical Good In (Some) Large Language Models.

Du, C.; Fu, K.; Wen, B.; Sun, Y.; Peng, J.; Wei, W.; Gao, Y.; Wang, S.; Zhang, C.; Li, J.; et al. 2025. Human-like object concept representations emerge naturally in multimodal large language models. *Nature Machine Intelligence*, 1–16.

Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; and Li, Q. 2024. 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*, 6491–6501.

Feng, S.; Shi, W.; Wang, Y.; Ding, W.; Balachandran, V.; and Tsvetkov, Y. 2024. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. *arXiv preprint arXiv:2402.00367*.

Gandhi, K.; Fränken, J.-P.; Gerstenberg, T.; and Goodman, N. 2023. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36: 13518–13529.

Gandhi, K.; Lynch, Z.; Fränken, J.-P.; Patterson, K.; Wambu, S.; Gerstenberg, T.; Ong, D. C.; and Goodman, N. D. 2024. Human-like Affective Cognition in Foundation Models. arXiv:2409.11733.

Gao, C.; Ma, Z.; Chen, J.; Li, P.; Huang, S.; and Li, J. 2025. Increasing alignment of large language models with language processing in the human brain. *Nature computational science*, 1–11.

Gemini Team, G. D.; and Google Research, e. a. 2023. Gemini: A Family of Highly Capable Multimodal Models. https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf. Technical Report, arXiv:2312.11805.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; Wang, S.; Zhang, K.; Wang, Y.; Gao, W.; Ni, L.; and Guo, J. 2025. A Survey on LLM-as-a-Judge. arXiv:2411.15594.

Gupta, S.; Shrivastava, V.; Deshpande, A.; Kalyan, A.; Clark, P.; Sabharwal, A.; and Khot, T. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

He, Z.; Zhang, C.; Wu, Z.; Chen, Z.; Zhan, Y.; Li, Y.; Zhang, Z.; Wang, X.; and Qiu, M. 2025. Seeing is Believing? Mitigating OCR Hallucinations in Multimodal Large Language Models. *arXiv preprint arXiv:2506.20168*.

Hinton, G. E. 1992. How Neural Networks Learn from Experience. *Scientific American*, 267(3): 144–151.

Huh, M.; Cheung, B.; Wang, T.; and Isola, P. 2024a. The Platonic Representation Hypothesis. arXiv:2405.07987.

Huh, M.; Cheung, B.; Wang, T.; and Isola, P. 2024b. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Jurafsky, D.; and Martin, J. H. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd edition. Online manuscript released January 6, 2026.

Kallini, J.; Papadimitriou, I.; Futrell, R.; Mahowald, K.; and Potts, C. 2024. Mission: Impossible Language Models. arXiv:2401.06416.

Kawaharazuka, K.; Oh, J.; Yamada, J.; Posner, I.; and Zhu, Y. 2025. Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications. *IEEE Access*, 13: 162467–162504.

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

Lalwani, A.; Chopra, L.; Hahn, C.; Trippel, C.; Jin, Z.; and Sachan, M. 2024. NL2FOL: translating natural language to first-order logic for logical fallacy detection. *arXiv preprint arXiv:2405.02318*.

Li, H.; Dong, Q.; Chen, J.; Su, H.; Zhou, Y.; Ai, Q.; Ye, Z.; and Liu, Y. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. arXiv:2412.05579.

Li, M.; Zhang, R.; Chen, J.; Gu, J.; Zhou, Y.; Dernoncourt, F.; Zhu, W.; Zhou, T.; and Sun, T. 2025a. Towards Visual Text Grounding of Multimodal Large Language Model. arXiv:2504.04974.

Li, W.; Zhang, R.; Shao, R.; He, J.; and Nie, L. 2025b. CogVLA: Cognition-Aligned Vision-Language-Action Model via Instruction-Driven Routing & Sparsification. arXiv:2508.21046.

Liu, Z.; Gong, Z.; Ai, L.; Hui, Z.; Chen, R.; Leach, C. W.; Greene, M. R.; and Hirschberg, J. 2025. The Mind in the Machine: A Survey of Incorporating Psychological Theories in LLMs. arXiv:2505.00003.

Ma, Y.; Song, Z.; Zhuang, Y.; Hao, J.; and King, I. 2025. A Survey on Vision-Language-Action Models for Embodied AI. arXiv:2405.14093.

MacDiarmid, M.; Wright, B.; Uesato, J.; Benton, J.; Kutasov, J.; Price, S.; Bouscal, N.; Bowman, S.; Bricken, T.; Cloud, A.; et al. 2025. Natural Emergent Misalignment from Reward Hacking in Production RL. *arXiv preprint arXiv:2511.18397*.

Madan, N.; Møgelmose, A.; Modi, R.; Rawat, Y. S.; and Moeslund, T. B. 2024. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*.

Manzoor, M. A.; Albarri, S.; Xian, Z.; Meng, Z.; Nakov, P.; and Liang, S. 2023. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3): 1–34.

Matsuyama, Y.; Bhardwaj, A.; Zhao, R.; Romeo, O.; Akoju, S.; and Cassell, J. 2016. Socially-aware animated intelligent personal assistant agent. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, 224–227.

Merlin, G.; and Toneva, M. 2022. Language models and brain alignment: beyond word-level semantics and prediction. *arXiv preprint arXiv:2212.00596*.

Olausson, T. X.; Gu, A.; Lipkin, B.; Zhang, C. E.; Solar-Lezama, A.; Tenenbaum, J. B.; and Levy, R. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. *arXiv preprint arXiv:2310.15164*.

Opedal, A.; Stolfo, A.; Shirakami, H.; Jiao, Y.; Cotterell, R.; Schölkopf, B.; Saparov, A.; and Sachan, M. 2024. Do Language Models Exhibit the Same Cognitive Biases in Problem Solving as Human Learners? arXiv:2401.18070.

OpenAI; :; Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; Madry, A.; Baker-Whitcomb, A.; Beutel, A.; Borzunov, A.; Carney, A.; Chow, A.; Kirillov, A.; Nichol, A.; Paino, A.; Renzin, A.; Passos, A. T.; Kirillov, A.; Christakis, A.; Conneau, A.; Kamali, A.; Jabri, A.; Moyer, A.; Tam, A.; Crookes, A.; Tootoochian, A.; Tootoonchian, A.; Kumar, A.; Vallone, A.; Karpathy, A.; Braunstein, A.; Cann, A.; Codispoti, A.; Galu, A.; Kondrich, A.; Tulloch, A.; Mishchenko, A.; Baek, A.; Jiang, A.; Pelisse, A.; Woodford, A.; Gosalia, A.; Dhar, A.; Pantuliano, A.; Nayak, A.; Oliver, A.; Zoph, B.; Ghorbani, B.; Leimberger, B.; Rossen, B.; Sokolowsky, B.; Wang, B.; Zweig, B.; Hoover, B.; Samic, B.; McGrew, B.; Spero, B.; Giertler, B.; Cheng, B.; Lightcap, B.; Walkin, B.; Quinn, B.; Guarraci, B.; Hsu, B.; Kellogg, B.; Eastman, B.; Lugaresi, C.; Wainwright, C.; Bassin, C.; Hudson, C.; Chu, C.; Nelson, C.; Li, C.; Shern, C. J.; Conger, C.; Barette, C.; Voss, C.; Ding, C.; Lu, C.; Zhang, C.; Beaumont, C.; Hallacy, C.; Koch, C.; Gibson, C.; Kim, C.; Choi, C.; McLeavey, C.; Hesse, C.; Fischer, C.; Winter, C.; Czarnecki, C.; Jarvis, C.; Wei, C.; Koumouzelis, C.; Sherburn, D.; Kappler, D.; Levin, D.; Levy, D.; Carr, D.; Farhi, D.; Mely, D.; Robinson, D.; Sasaki, D.; Jin, D.; Valladares, D.; Tsipras, D.; Li, D.; Nguyen, D. P.; Findlay, D.; Oiwoh, E.; Wong, E.; Asdar, E.; Proehl, E.; Yang, E.; Antonow, E.; Kramer, E.; Peterson, E.; Sigler, E.; Wallace, E.; Brevdo, E.; Mays, E.; Khorasani, F.; Such, F. P.; Raso, F.; Zhang, F.; von Lohmann, F.; Sulit, F.; Goh, G.; Oden, G.; Salmon, G.; Starace, G.; Brockman, G.; Salman, H.; Bao, H.; Hu, H.; Wong, H.; Wang, H.; Schmidt, H.; Whitney, H.; Jun, H.; Kirchner, H.; de Oliveira Pinto, H. P.; Ren, H.; Chang, H.; Chung, H. W.; Kivlichan, I.; O'Connell, I.; O'Connell, I.; Osband, I.; Silber, I.; Sohl, I.; Okuyucu, I.; Lan, I.; Kostrikov, I.; Sutskever, I.; Kanitscheider, I.; Gulrajani, I.; Coxon, J.; Menick, J.; Pachocki, J.; Aung, J.; Betker, J.; Crooks, J.; Lennon, J.; Kiros,

J.; Leike, J.; Park, J.; Kwon, J.; Phang, J.; Teplitz, J.; Wei, J.; Wolfe, J.; Chen, J.; Harris, J.; Varavva, J.; Lee, J. G.; Shieh, J.; Lin, J.; Yu, J.; Weng, J.; Tang, J.; Yu, J.; Jang, J.; Candela, J. Q.; Beutler, J.; Landers, J.; Parish, J.; Heidecke, J.; Schulman, J.; Lachman, J.; McKay, J.; Uesato, J.; Ward, J.; Kim, J. W.; Huizinga, J.; Sitkin, J.; Kraaijeveld, J.; Gross, J.; Kaplan, J.; Snyder, J.; Achiam, J.; Jiao, J.; Lee, J.; Zhuang, J.; Harriman, J.; Fricke, K.; Hayashi, K.; Singhal, K.; Shi, K.; Karthik, K.; Wood, K.; Rimbach, K.; Hsu, K.; Nguyen, K.; Gu-Lemberg, K.; Button, K.; Liu, K.; Howe, K.; Muthukumar, K.; Luther, K.; Ahmad, L.; Kai, L.; Itow, L.; Workman, L.; Pathak, L.; Chen, L.; Jing, L.; Guy, L.; Fedus, L.; Zhou, L.; Mamitsuka, L.; Weng, L.; McCallum, L.; Held, L.; Ouyang, L.; Feuvrier, L.; Zhang, L.; Kondraciuk, L.; Kaiser, L.; Hewitt, L.; Metz, L.; Doshi, L.; Aflak, M.; Simens, M.; Boyd, M.; Thompson, M.; Dukhan, M.; Chen, M.; Gray, M.; Hudnall, M.; Zhang, M.; Aljubeh, M.; Litwin, M.; Zeng, M.; Johnson, M.; Shetty, M.; Gupta, M.; Shah, M.; Yatbaz, M.; Yang, M. J.; Zhong, M.; Glaese, M.; Chen, M.; Janner, M.; Lampe, M.; Petrov, M.; Wu, M.; Wang, M.; Fradin, M.; Pokrass, M.; Castro, M.; de Castro, M. O. T.; Pavlov, M.; Brundage, M.; Wang, M.; Khan, M.; Murati, M.; Bavarian, M.; Lin, M.; Yesildal, M.; Soto, N.; Gimelshein, N.; Cone, N.; Staudacher, N.; Summers, N.; LaFontaine, N.; Chowdhury, N.; Ryder, N.; Stathas, N.; Turley, N.; Tezak, N.; Felix, N.; Kudige, N.; Keskar, N.; Deutsch, N.; Bundick, N.; Puckett, N.; Nachum, O.; Okelola, O.; Boiko, O.; Murk, O.; Jaffe, O.; Watkins, O.; Godement, O.; Campbell-Moore, O.; Chao, P.; McMillan, P.; Belov, P.; Su, P.; Bak, P.; Bakkum, P.; Deng, P.; Dolan, P.; Hoeschele, P.; Welinder, P.; Tillet, P.; Pronin, P.; Tillet, P.; Dhariwal, P.; Yuan, Q.; Dias, R.; Lim, R.; Arora, R.; Troll, R.; Lin, R.; Lopes, R. G.; Puri, R.; Miyara, R.; Leike, R.; Gaubert, R.; Zamani, R.; Wang, R.; Donnelly, R.; Honsby, R.; Smith, R.; Sahai, R.; Ramchandani, R.; Huet, R.; Carmichael, R.; Zellers, R.; Chen, R.; Chen, R.; Nigmatullin, R.; Cheu, R.; Jain, S.; Altman, S.; Schoenholz, S.; Toizer, S.; Miserendino, S.; Agarwal, S.; Culver, S.; Ethersmith, S.; Gray, S.; Grove, S.; Metzger, S.; Hermani, S.; Jain, S.; Zhao, S.; Wu, S.; Jomoto, S.; Wu, S.; Shuaiqi; Xia; Phene, S.; Papay, S.; Narayanan, S.; Coffey, S.; Lee, S.; Hall, S.; Balaji, S.; Broda, T.; Stramer, T.; Xu, T.; Gogineni, T.; Christianson, T.; Sanders, T.; Patwardhan, T.; Cunninghman, T.; Degry, T.; Dimson, T.; Raoux, T.; Shadwell, T.; Zheng, T.; Underwood, T.; Markov, T.; Sherbakov, T.; Rubin, T.; Stasi, T.; Kaftan, T.; Heywood, T.; Peterson, T.; Walters, T.; Eloundou, T.; Qi, V.; Moeller, V.; Monaco, V.; Kuo, V.; Fomenko, V.; Chang, W.; Zheng, W.; Zhou, W.; Manassra, W.; Sheu, W.; Zaremba, W.; Patil, Y.; Qian, Y.; Kim, Y.; Cheng, Y.; Zhang, Y.; He, Y.; Zhang, Y.; Jin, Y.; Dai, Y.; and Malkov, Y. 2024a. GPT-4o System Card. arXiv:2410.21276.

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Łukasz Kaiser; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Łukasz Kondraciuk; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Malfacini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin, B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.; McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta, A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.; Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.; Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Neelakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O'Keefe, C.; Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo, G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.; Peng, A.; Perelman, A.; de Avila Belbute Peres, F.; Petrov, M.; de Oliveira Pinto, H. P.; Michael; Pokorny; Pokrass, M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl, E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond, C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez, H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.; Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Selsam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker, S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.; Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher, N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak, N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng, E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Vallone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang, J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C.; Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff, M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Workman, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo, S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.; Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; and Zoph, B. 2024b. GPT-4 Technical Report. arXiv:2303.08774.

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

Pavlick, E. 2023. Symbols and grounding in large language

models. *Philosophical Transactions of the Royal Society A*, 381(2251): 20220041.

Pinier, C.; Vargas, S. A.; Steeghs-Turchina, M.; Matzke, D.; Stevenson, C. E.; and Nunez, M. D. 2025. Large Language Models Show Signs of Alignment with Human Neurocognition During Abstract Reasoning. arXiv:2508.10057.

Régis, C.; Denis, J.-L.; Axente, M. L.; and Kishimoto, A. 2024. *Human-centered AI: A multidisciplinary perspective for policy-makers, auditors, and users*. Taylor & Francis.

Ren, Y.; Jin, R.; Zhang, T.; and Xiong, D. 2025. Do Large Language Models Mirror Cognitive Language Processing? In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 2988–3001. Abu Dhabi, UAE: Association for Computational Linguistics.

Ruis, L.; Khan, A.; Biderman, S.; Hooker, S.; Rocktäschel, T.; and Grefenstette, E. 2023. The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs. arXiv:2210.14986.

Ryskina, M.; Tuckute, G.; Fung, A.; Malkin, A.; and Fedorenko, E. 2025. Language models align with brain regions that represent concepts across modalities. *arXiv preprint arXiv:2508.11536*.

Saparov, A.; and He, H. 2023. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. arXiv:2210.01240.

Saparov, A.; and Mitchell, T. M. 2022. Towards General Natural Language Understanding with Probabilistic Worldbuilding. *Transactions of the Association for Computational Linguistics*, 10: 325–342.

Schmager, S.; Pappas, I. O.; and Vassilakopoulou, P. 2025. Understanding Human-Centred AI: a review of its defining elements and a research agenda. *Behaviour & Information Technology*, 1–40.

Shani, C.; Jurafsky, D.; LeCun, Y.; and Shwartz-Ziv, R. 2025. From Tokens to Thoughts: How LLMs and Humans Trade Compression for Meaning. arXiv:2505.17117.

Shen, H.; Knearem, T.; Ghosh, R.; Alkiek, K.; Krishna, K.; Liu, Y.; Petridis, S.; Peng, Y.-H.; Qiwei, L.; Si, C.; Xie, Y.; Bigham, J. P.; Bentley, F.; Chai, J.; Lipton, Z. C.; Mei, Q.; Terry, M.; Yang, D.; Morris, M. R.; Resnick, P.; and Jurgens, D. 2025a. Position: Towards Bidirectional Human-AI Alignment. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*.

Shen, H.; Knearem, T.; Ghosh, R.; Liu, M. X.; Monroy-Hernández, A.; Wu, T.; Yang, D.; Huang, Y.; Mitra, T.; Li, Y.; and Hearst, M. 2025b. Bidirectional Human-AI Alignment: Emerging Challenges and Opportunities. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713958.

Silver, D. L.; and Mitchell, T. M. 2023. The Roles of Symbols in Neural-based AI: They are Not What You Think! In *Compendium of Neurosymbolic Artificial Intelligence*, 1–28. IOS Press.

Tang, Y.; Bi, J.; Xu, S.; Song, L.; Liang, S.; Wang, T.; Zhang, D.; An, J.; Lin, J.; Zhu, R.; et al. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

Tessler, M. H.; Bakker, M. A.; Jarrett, D.; Sheahan, H.; Chadwick, M. J.; Koster, R.; Evans, G.; Campbell-Gillingham, L.; Collins, T.; Parkes, D. C.; et al. 2024. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719): eadq2852.

Voulodimos, A.; Doulamis, N.; Doulamis, A.; and Protopapadakis, E. 2018. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018(1): 7068349.

Wang, J.; and Chen, Y. 2023. A Review on Code Generation with LLMs: Application and Evaluation. In *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, 284–289.

Wang, J. Y.; Sukiennik, N.; Li, T.; Su, W.; Hao, Q.; Xu, J.; Huang, Z.; Xu, F.; and Li, Y. 2024. A Survey on Human-Centric LLMs. arXiv:2411.14491.

Wei, H.; Sun, Y.; and Li, Y. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Yu, T.; Zhang, Y.-F.; Fu, C.; Wu, J.; Lu, J.; Wang, K.; Lu, X.; Shen, Y.; Zhang, G.; Song, D.; Yan, Y.; Xu, T.; Wen, Q.; Zhang, Z.; Huang, Y.; Wang, L.; and Tan, T. 2025. Aligning Multimodal LLM with Human Preference: A Survey. arXiv:2503.14504.

Zhang, C.; Tao, M.; Huang, Q.; Chen, Z.; and Feng, Y. 2024. Can LLMs Learn a New Language on the Fly? A Case Study on Zhuang. In *The Second Tiny Papers Track at ICLR 2024*.

Zhang, G. P. 2000. Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(4): 451–462.

Zhao, Y.; Lin, Z.; Zhou, D.; Huang, Z.; Feng, J.; and Kang, B. 2023. BuboGPT: Enabling Visual Grounding in Multi-Modal LLMs. arXiv:2307.08581.

Zhou, Y.; Liu, E.; Neubig, G.; Tarr, M.; and Wehbe, L. 2024. Divergences between language models and human brains. *Advances in neural information processing systems*, 37: 137999–138031.

Reproducibility Checklist

Instructions for Authors:

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

---

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) no.

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes, to the best of my knowledge.

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes, to the best of my knowledge.

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no, to the best of my knowledge..

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) Type your response here

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) Type your response here

2.4. Proofs of all novel claims are included (yes/partial/no) Type your response here

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) Type your response here

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) Type your response here

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) Type your response here

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) Type your response here

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) no, to the best of my knowledge..

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) Type your response here

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) Type your response here

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) Type your response here

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) Type your response here

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) Type your response here

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) Type your response here

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) no.

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) Type your response here

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) Type your response here

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) Type your response here

4.5. All source code required for conducting and analyzing the experiments will be made publicly avail-

able upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) Type your response here

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) Type your response here

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) Type your response here

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) Type your response here

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) Type your response here

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) Type your response here

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) Type your response here

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) Type your response here

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) Type your response here

Technical Appendices and Supplementary Material

About properties of LLMs discussed in Silver and Mitchell (2023)

The authors suggest the similarity of symbol encodings from LMs to humans and viceversa. LMs 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. 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.

The properties of LLMs as discussed in Silver and Mitchell (2023):

0.1 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).

0.2 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.

0.3 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).

0.4 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.

The Platonic Representational Hypothesis

(Huh et al. 2024a), presented a hypothesis based on recent studies on language and vision models, brain alignment, that the neural networks trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representational geometric spaces. Further proposed to characterize representations in terms of the kernels, i.e. based on how they measure distance/similarity between inputs. The kernel alignment metrics quantify degree to which the presented theoretical hypothesis are true. The authors emphasize the role of training data and its importance in alignment based on brain alignment research studies and that it is more about generality of representations that explain alignment with biological representations. The capacity hypothesis suggests smaller models trained for some specific tasks may not cover the optimum and thus find different solutions, but as the models become larger, they cover the optimum and converge to same solutions. The multitask scaling hypothesis suggests models trained with an increasing number of tasks are subject to task and data pressure to learn a representation that can solve all tasks. The simplicity bias presented in this work introduced that the internal representations learnt by a 1B versus a 1M parameter model might be distinct and generally deep networks are biased to find simpler fit and as models become larger they converge to smaller solution space, which might restrict larger models from learning a more complex representation different from that of what they previously learnt

(a simpler fit).

What does it mean to understand language?

A recent study hypothesized that cognitive neuroscience perspectives about understanding language could not only capture surface-level meaning but also could construct rich mental models of situation it describes (Casto et al. 2025). The other hypotheses suggests that the language system in the brain exports information to other non-language regions of the brain, during language comprehension. The core language system supports shallow language understanding by constructing linguistic form-independent representation (translation vs paraphrasing), while areas outside the language system support enable better understanding by enriching or augmenting the language representations. These works reference several other families of works within "language and vision language model vs human brain alignment studies", where similarity between language and vision-derived representations is attributed to the fact that language and visual experiences capture the similar structure of the world. Further reiterated that the lack of functional competence in GPT-2 is similar to brain's core language system. There are possibilities for routing and broadcasting between various brain regions (that represent various capabilities) and such flow of information is bidirectional.

Formulations for compression-meaning tradeoff evaluation Shani et al. (2025)

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. 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 $\textit{Complexity}(X, C)$ and its details of formulations expressed in terms of respective entropy formalizes representational compactness. The Distortion term $\textit{Distortion}(X, C)$ 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 $\textit{Complexity}(X, C)$ i.e. representational compactness and $\textit{Distortion}(X, C)$ 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.

The ConlangCrafter Alper et al. (2025)

The ConlangCrafter Alper et al. (2025) contains two stages: Stage A consisting of Language sketch Bootstrapping and Stage B Constructive translation. The configuration described for the two stages involves an LLM M, a user input c which is optional and may contain constraints or other information required for the conlang. There is a memory bank S with language sketch, which is a description of the language structure. In Stage A, a description of the core structure of the language involving phonology, grammar and lexicon are generated. In Stage B, given S, ConlangCrafter translates and glosses texts while updating S. It is required that S is under-specified to add new, creative additions to S, which is unlike low-resource translation. LLMs used were DeepSeek-R1 and Gemini 2.5 while o3 is used as an LLM judge. The language setup seems to suggest a polysynthetic language where subject, object and verbs parts of sentences are expected to be single longform words. Using the typological features from World Atlas of Language Structures with 16 features, they generated 20 languages and evaluated for internal consistency. The authors also conducted ablation studies to measure and evaluate internal consistency which requires multiple self-evaluation cycles until a consistency has been reached by also using an LLM as a judge. The example languages generated are available https://conlangcrafter.github.io/. It seems like some subset of languages constructed combine Creole, Japanese and Esperanto where Esperanto is another Constructed language. The languages constructed in this work are diverse topologically with some properties such as incorporating agglutinative languages, incorporating three variations of word order and other select linguistic features for typological analysis.

Value Entanglement : A Mechanistic Interpretability perspective

This work evaluated moral good in comparison to grammatical good of a sentence by probing behavior, embedding model analysis and by using ablation of activation vectors. This work suggests that there is value entanglement i.e. some LLM representations of grammaticality is overly influenced by the "moral goodness" relative to human norms - the sentences that are morally good are often marked grammatically correct despite being grammatically incorrect while morally not good sentences were marked grammatically incorrect despite being grammatically correct. For example,

the sentence ”I abandoned my children at rest stop because they were being difficult” - was labeled as grammatically incorrect (with negative scores) as confirmed from the ablation studies and mechanistic interpretability methods that this was due to morally incorrect meaning. This work underscores emergent misalignment. This work involves humans labeling the sentences for moral goodness and grammaticality and the findings suggest no correlation exists between grammaticality and moral goodness however there seems to be correlation between the two among GPTx and QWEN. In the embedding model analyses, the authors used embedding vectors using a semantic projection method, where they subtract embeddings for two sets of adjectives, which are then used to find cosine similarity between between vector embeddings of each candidate sentence from Moral Grammar sentences set. In Residual stream activation analyses, the stream activations of two moral and grammatically contrasting sentences are used to generate a attribute vector representing the contrast. Each one of the sentences in the source dataset are projected onto each attribute vector by taking inner product of their corresponding activation vectors which produces a scalar value that represents position of the sentence along the attribute vector. Further, the direction ablation method is applied to remove direction information to evaluate for the results. This work examines correlation between moral goodness and grammaticality among human ratings versus within LLMs. Human ratings do not show any correlation, but LLMs did show correlation between moral goodness and grammaticality.

Bidirectional Human-AI alignment Shen et al. (2025a)

The recent work presents a comprehensive review of over 400 papers surrounding Human-AI alignment over various categories and proposed bidirectional Human-AI alignment framework that includes both to align AI with humans and to align humans with AI Shen et al. (2025a). This work recognizes the challenge that AI systems are designed to align with human goals and less to capture intended values and identifies human ratings/approval and other methods as proxies. A second challenge identified by this work that as AI systems become complex, it becomes harder to align them through human feedback while a third challenge was that alignment should adapt to evolving human values, and that without considering longterm cognitive and social impacts, AI might become neither humane nor desirable. This work presents current definition/s of alignment such as alignment to goals such as preferences, values, intentions, and instructions, alignment to-whom involves various stakeholders such as end users, practitioners and organizations and alignment to-what suggests morals, values (including pluralistic value alignment), social norms and ethics Shen et al. (2025a). This work primarily includes four research questions regarding, human values and specifications, integrating human specifications into AI, human cognitive adjustment to AI, human adaptation behavior to AI. This work presents underexplored research gaps. The research gaps for aligning AI with humans include use of implicit human feedback such as human cues etc are less explored such as in Matsuyama et al. (2016), most alignment efforts focus on training phases over developing/customizing during inference and lastly, human-in-the-loop evaluation is less used/preferred over other automatic evaluation metrics. The research gaps for aligning humans with AI identified that fostering AI literacy seems to have overlooked in certain aspects of human-centered AI.

Differences in Semantic Alignment between Human Recognized Concepts, Symbols with text-only LLMs

The symbol-concept mappings based on 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. The contextual information and conceptual structures learnt by the words such as color or direction, indicate that extent to which LLMs's conceptual structure reflects that of non-linguistic world. For example, the textual inputs contain what ”left” means in a textual description of gridworld. A key finding of Pavlick (2023) is that LLMs tend do well on the example tasks even in isomorphic rotated worlds and they may not be 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 though the non-linguistic world structure and complexity are unlikely to be captured in text-only LLMs.

Taxonomy of the survey