Knowledge Mechanisms in Large Language Models: A Survey and Perspective

Mengru Wang[1]*, Yunzhi Yao[1]*, Ziwen Xu[1], Shuofei Qiao[1], Shumin Deng[2], Peng Wang[1], Xiang Chen[1], Jia-Chen Gu[3], Yong Jiang[4], Pengjun Xie[4], Fei Huang[4], Huajun Chen[1], Ningyu Zhang[1]†

[1]Zhejiang University, [2]National University of Singapore, NUS-NCS Joint Lab, Singapore, [3]University of California, Los Angeles, [4]Alibaba Group

{mengruwg,zhangningyu}@zju.edu.cn

Abstract

Understanding knowledge mechanisms in Large Language Models (LLMs) is crucial for advancing towards trustworthy AGI. This paper reviews knowledge mechanism analysis from a novel taxonomy including knowledge utilization and evolution. Knowledge utilization delves into the mechanism of memorization, comprehension and application, and creation. Knowledge evolution focuses on the dynamic progression of knowledge within individual and group LLMs. Moreover, we discuss what knowledge LLMs have learned, the reasons for the fragility of parametric knowledge, and the potential dark knowledge (hypothesis) that will be challenging to address. We hope this work can help understand knowledge in LLMs and provide insights for future research.

1 Introduction

Knowledge is the cornerstone of intelligence and the continuation of civilization, furnishing us with foundational principles and guidance for navigating complex problems and emerging challenges (Davis et al., 1993; Choi, 2022). Throughout the extensive history of evolution, we have dedicated our lives to cultivating more advanced intelligence by utilizing acquired knowledge and exploring the frontiers of unknown knowledge (McGraw and Harbison-Briggs, 1990; Han et al., 2021).

As we know, Large language models (LLMs) are renowned for encapsulating extensive parametric knowledge (Roberts et al., 2020; Sung et al., 2021; Cao et al., 2021a; Zhong et al., 2021; Kandpal et al., 2023; Heinzerling and Inui, 2020; Petroni et al., 2019; Qiao et al., 2023; Kritharoula et al., 2023; He et al., 2024a), achieving unprecedented progress in application. However, *the knowledge mechanisms in LLMs for learning, storage, utilization, and evolution still remain mysterious* (Phillips et al., 2021; Gould et al., 2023a). Extensive works aim to demystify various types of knowledge in LLMs through knowledge neurons (Dai et al., 2022; Chen et al., 2024a) and circuits (Elhage et al., 2021; Yao et al., 2024; Zou et al., 2024), yet these efforts, scattered across various tasks, await comprehensive review and analysis.

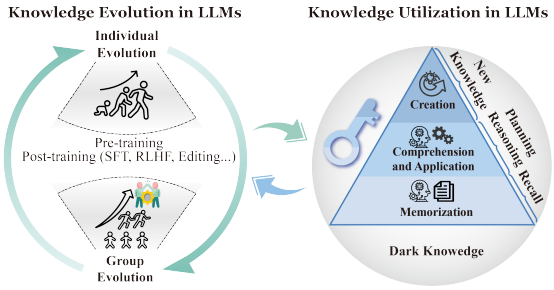

Figure 1: The analysis framework of knowledge mechanism within neural models includes knowledge evolution and utilization. Dark knowledge denotes knowledge unknown to human or model (machine). We investigate the mechanisms of knowledge utilization (right) in LLMs during a specific period of their evolution (left). The knowledge limitations identified through mechanisms analysis will inspire subsequent evolution (left).

This paper pioneeringly reviews the mechanism across the whole knowledge life cycle (as shown in Fig 1). We also propose a novel taxonomy for knowledge mechanisms in LLMs, as illustrated in Fig 4, which encompasses knowledge utilization at a specific time and knowledge evolution across all periods of LLMs [1]. Specifically, we introduce preliminaries of this field (§A) and review the knowledge utilization mechanism from a new perspective (§2), delve into the fundamental principles for knowledge evolution (§3). Then, we investigate how to construct more efficient and trustworthy LLMs from the perspective of knowledge mechanism (§E). Later, We discuss open questions about

*Equal Contribution.
†Corresponding Author.

[1]Knowledge utilization focuses on *static* knowledge at a specific period, while knowledge evolution explores the long-term *dynamic* development of knowledge across individual and group LLMs.

Findings of the Association for Computational Linguistics: EACL 2024, pages 7097–7135
November 12-16, 2024

the knowledge LLMs have and have not acquired (§4). Finally, we also provide some future directions (§G) and tools for knowledge mechanism analysis (§D). Our contributions are as follows:

- To the best of our knowledge, we are the first to review knowledge mechanisms in LLMs and provide **a novel taxonomy** across the entire life.
- We propose a new perspective to analyze knowledge utilization mechanisms from **three levels: memorization, comprehension and application, and creation**.
- We discuss **knowledge evolution** in individual and group LLMs, and analyze the inherent conflicts and integration in this process.
- We observe that LLMs have learned basic world knowledge. However, the learned knowledge is fragile, leading to challenges such as **hallucinations and knowledge conflicts**. We speculate that this fragility may be primarily due to improper learning data. Besides, the unlearned **dark knowledge** will exist long.

Comparison with Existing Surveys Previous interpretability surveys typically aim to investigate various *methods for explaining the roles of different components* within LLMs from the global and local taxonomy (Ferrando et al., 2024; Zhao et al., 2024a; Luo and Specia, 2024; Murdoch et al., 2019; Rai et al., 2024a; Bereska and Gavves, 2024; Vilas et al., 2024; Singh et al., 2024). In contrast, this paper focuses on knowledge in LLMs. Hence, *our taxonomy, oriented from target knowledge in LLMs, reviews how knowledge is acquired, stored, utilized, and subsequently evolves*. Additionally, previous taxonomy mostly explore the explainability *during the inference stage* (a specific period), while ignoring knowledge acquisition during the pre-training stage and evolution during the post-training stage (Räuker et al., 2023; Luo et al., 2024b; Apidianaki, 2023; Jiao et al., 2023; Räuker et al., 2023; Rai et al., 2024b). Our taxonomy aims to explore the *dynamic evolution across all periods* from naivety to sophistication in both individual and group LLMs. In contrast to the most similar survey (Cao et al., 2024a) that introduces knowledge life cycle, our work focuses on the underlying mechanisms at each stage.

Generally, this paper may help us to explore and manipulate advanced knowledge in LLMs, examine current limitations through the history of knowledge evolution, and **inspire more efficient and trustworthy architecture and learning strategy for future models from knowledge mechanism perspective**. Note that most hypotheses in this paper are derived from transformer-based LLMs. We also validate the generalizability of these hypotheses across other architectural models and then propose universality intelligence in §C.

2 Knowledge Utilization in LLMs

Knowledge is an awareness of facts, a form of familiarity, awareness, understanding, or acquaintance (Zagzebski, 2017; Hyman, 1999; Mahowald et al., 2023; Gray et al., 2024). It often involves the possession of information learned through experience and can be understood as a cognitive success or an epistemic contact with reality. We also introduce some preliminary information in §A, which includes the definition of knowledge in LLMs and knowledge analysis methods.

Then, inspried by Bloom's Taxonomy of cognition levels (Wilson, 2016; Bloom et al., 1956; Keene et al., 2010; Fadul, 2009), we categorize knowledge representation and utilization within LLMs into three levels (as shown in Fig 2): memorization, comprehension and application, and creation [2]. Note that these mechanistic analyses are implemented via methods in §A.4. We further evaluate the applicability, advantages, and limitations of different methods in §B.1.

2.1 Memorization

Knowledge memorization (Schwarzschild et al., 2024; Prashanth et al., 2024) aims to remember and recall knowledge in the training corpus, e.g., specific terms (entities), grammar, facts, commonsense, concepts, etc (Allen-Zhu and Li, 2023a; Yu et al., 2023a; Mahowald et al., 2023; Zhu and Li, 2023; Allen-Zhu and Li, 2023b, 2024; Cao et al., 2024a). *We posit knowledge memorization from Modular Region and Connection Hypothesis by reviewing existing research.*

Hypothesis 1: Modular Region

Knowledge is Encoded in Modular Regions.

[2]Note that we combine analyzing, evaluating, and creating from Bloom's Taxonomy into one category level (creation) in our taxonomy, as they are difficult to disentangle. Specifically, creation emphasizes the capacity and process of forming *novel* and *valuable* things. Analyzing (Wilson, 2016), which breaks materials or concepts into parts, is used for creating *novel* things. Evaluating (Wilson, 2016) is usually used for assessing the *value* of new creations.

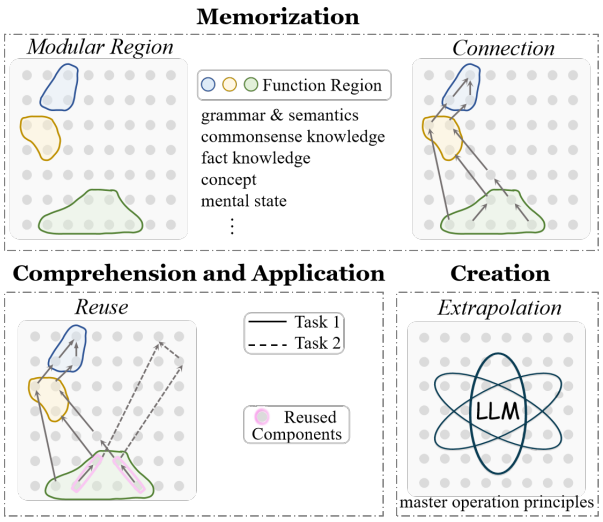

Figure 2: The mechanism analysis for knowledge utilization across three levels: memorization, comprehension and application, and creation.

This modular region hypothesis simplifies knowledge representation in transformer-based models into isolated modular region, e.g., MLPs or attention heads. **Knowledge is encoded via MLPs.** Geva et al. (2021) posit that MLPs operate as key-value memories and each individual key vector corresponds to a specific *semantic pattern* or *grammar*. Based on the above finding, Geva et al. (2022b,a) reverse engineer the operation of the MLPs layers and find that MLPs can promote both semantic (e.g., measurement semantic including kg, percent, spread, total, yards, pounds, and hours) and *syntactic* (e.g., adverbs syntactic including largely, rapidly, effectively, previously, and normally) concepts in the vocabulary space. Miller and Neo (2024) find a single MLP neuron (in GPT-2 Large) capable of generating "an" or "a". Subsequently, *fact* (Dai et al., 2022; Meng et al., 2022) and *commonsense* knowledge (Gupta et al., 2023) are found. Advanced language-specific neurons (Tang et al., 2024), linguistic regions (Zhao et al., 2023a), entropy neurons (Stolfo et al., 2024), abstract conceptual (Wang et al., 2024e) and unsafe (Wang et al., 2024b; Wu et al., 2023a) knowledge, are also observed in MLPs. In addition to MLP, **knowledge is also conveyed by attention heads** (Geva et al., 2023; Gould et al., 2023b). Hoover et al. (2020) explain the knowledge each attention head has learned. Specifically, attention heads store evident *linguistic features, positional information*, and so on. Besides, *fact knowledge* (Yu et al., 2023c; Li et al., 2023a) and *bias* (Hoover et al., 2020) are mainly convey by attention heads. Jiang et al. (2024b) further observe that LLMs leverage self-attention to gather information through certain tokens in the contexts, which serve as clues, and use the value matrix for associative memory. Later, Zhu et al. (2024) also find that attention heads can simulate *mental state* and activate "Theory of Mind" (ToM) capability.

However, Hypothesis 1 ignores the connections between different regions. Inspired by advancements in neuroscience (de Schotten et al., 2022), Hypothesis 2 asserts that the connection of different components integrates knowledge, rather than the isolated regions in Hypothesis 1.

Hypothesis 2: Connection

Knowledge is Represented by Connections.

Geva et al. (2023) outline the encoding of factual knowledge (e.g., "The capital of Ireland is Dublin") through the following three steps: (1) subject (Ireland) information enrichment in MLPs, (2) the relation (capital of) propagates to the last token, (3) object (Dublin) is extracted by attention heads in later layers. This claim is supported by Li et al. (2024d). Similarly, Lv et al. (2024) conclude that task-specific attention head may move the topic entity to the final position of the residual stream, while MLPs conduct relation function. Moreover, the recent prominent knowledge circuit framework (Nainani, 2024; Yao et al., 2024; He et al., 2024b; Elhage et al., 2021; Marks et al., 2024) advocates leveraging a critical computational subgraph among all components to explore internal knowledge within LLM parameters. The competencies for indirect object identification and color object tasks are discovered to be embedded in specialized knowledge circuits (Conmy et al., 2023; Wang et al., 2023c; Merullo et al., 2023a; Yu et al., 2024c). Lan et al. (2024) also identify number-related circuits that encode the predictive ability of Arabic numerals, number words, and months. More importantly, experimental evidence demonstrates that various types of knowledge, including linguistic, commonsense, factual, and biased information, are encapsulated in specific knowledge circuits (Yao et al., 2024). Interestingly, knowledge encoded by specific circuits can rival or even surpass that of the entire LLM. This may be because knowledge circuits memorized the relevant knowledge, while noise from other components might impede the model's performance on these tasks.

2.2 Comprehension and Application

Knowledge comprehension and application focus on demonstrating the understanding of memorized knowledge and then solving problems in new situations, e.g., *generalization on out-of-domain tasks* (Wang et al., 2024a), *reasoning (Hou et al., 2023) and planning* (McGrath et al., 2021). Merrill et al. (2023) denote the transition from memorization to comprehension and application as grokking, and suggest that the grokking derives from two largely distinct subnetworks competition. Intuitively, only knowledge that is correctly memorized (Prashanth et al., 2024) in §2.1 can be further applied to solving complex tasks. Therefore, we posit the following Reuse Hypothesis from two knowledge memorization perspectives.

Hypothesis 3: Reuse

LLMs Reuse Certain Components during Knowledge Comprehension and Application.

From the Modular Region Perspective, knowledge utilization reuses some regions. These regions might include a few neurons, attention heads, MLPs, a transformer layer, or partial knowledge circuits. Generally, basic knowledge (position information, n-gram pattern, syntactic features) tends to be stored at earlier layers, while sophisticated knowledge (mental state, emotion, and abstract concept, e.g., prime number, Camelidae, and safety) is located at later layers (Zhu et al., 2024; Jin et al., 2024a; Wang et al., 2024b,e; Men et al., 2024; Kobayashi et al., 2023). Therefore, *neurons of earlier layers related to basic knowledge tend to be reused* (Kang and Choi, 2023; Zhao et al., 2024a; Kandpal et al., 2023). Various math reasoning tasks also utilize the attention mechanism in initial layers to map input information to the final token positions, subsequently generating answers using a set of MLPs in later layers (Stolfo et al., 2023; Hanna et al., 2023; Langedijk et al., 2023). Besides, *some specific function regions are also reused*. Specifically, retrieval heads (Li et al., 2023a) are reused for Chain-of-Thought (CoT) reasoning and long-context tasks. These retrieval heads are found in 4 model families, 6 model scales, and 3 types of fine-tuning. Subsequently, induction heads, identified in Llama and GPT, are claimed to be reused for in-context learning (ICL) tasks Olsson et al. (2022); Crosbie and Shutova (2024). Attention heads can map country names to their capitals in capital city-related tasks (Lv et al., 2024). Language-specific neurons (in Llama and BLOOM) are responsible for multiple language related tasks, such as English, French, Mandarin, and others Tang et al. (2024). Zhao et al. (2023a) further reveal linguistic regions (in Llama) correspond to linguistic competence, which is the cornerstone for performing various tasks. Later, function regions related to the process of math reasoning are also discovered in LLMs. For instance, the last layer of GPT-2 (trained from scratch) has been observed to exhibit mathematical reasoning abilities across various math questions (Ye et al., 2024). **From the Connection Perspective, knowledge utilization shares partial knowledge circuits.** For instance, similar tasks share subgraphs (computational circuits) with analogous roles (Lan et al., 2024). Besides, knowledge circuits (in GPT2) are reused to solve a seemingly different task, e.g., indirect object identification and colored objects tasks (Merullo et al., 2023a). Wang et al. (2024a) further observe that two-hop composition reasoning tasks reuse the knowledge circuits from the first hop. Yao et al. (2024) also believe that this reuse phenomenon exists in factual recall and multi-hop reasoning. Specifically, sub-circuits are reused in similar factual knowledge, such as tasks related to "city_in_country", "name_birth_place", and "country_language". Besides, Dutta et al. (2024) demystify LLMs how to perform CoT reasoning, i.e., Llama facilitates CoT tasks via multiple parallel circuits enjoying significant intersection.

2.3 Creation

Knowledge creation (Runco and Jaeger, 2012; Sternberg, 2006) emphasizes the capacity and process of forming *novel* and *valuable* things, rather than the existing ones (i.e., LLMs have seen) discussed in §2.1 and §2.2. The creations encompass two levels: 1) LLMs create new terms following the current world's principles comprehended by LLMs, such as new proteins (Shin et al., 2021), molecules (Bagal et al., 2022; Fang et al., 2023; Edwards et al., 2022), code (DeLorenzo et al., 2024), video (Kondratyuk et al., 2023), models (Zheng et al., 2024), names for people and companies, written stories (Pépin et al., 2024; Gómez-Rodríguez and Williams, 2023; Buz et al., 2024), synthetic data (Stenger et al., 2024; Mumuni et al., 2024; Abufadda and Mansour, 2021), etc. These novel items operate according to the existing rules, e.g., law

of conservation of energy, reasoning logic (Wang et al., 2024a), or principles of probability theory. 2) LLMs may generate new rules, such as mathematical theorems, and the resulting terms will operate according to the new rules. We posit that the knowledge creation of LLMs may derive from the Extrapolation Hypothesis.

Hypothesis 4: Extrapolation

LLMs May Create Knowledge via Extrapolation.

The expression of knowledge is diverse; some knowledge is inherently continuous. Therefore, it is difficult, if not impossible, to represent certain knowledge using discrete data points (Spivey and Michael, 2007; Penrose; Markman, 2013). LLMs utilize insights into the operational principles of the world to extrapolate additional knowledge from known discrete points, bridging gaps in knowledge and expanding our understanding of the world(Heilman et al., 2003; Douglas et al., 2024; Park et al., 2023b; Kondratyuk et al., 2023). Drawing inspiration from research on human creativity (Haase and Hanel, 2023), the *physical implementation of knowledge extrapolation relies on the plasticity of neurons* (Mukherjee and Chang, 2024). Specifically, plasticity refers to LLMs changing activations and connectivity between neurons according to the input (Coronel-Oliveros et al., 2024).

However, from statistical perspective, the intricate connections and activations between neurons, though not infinite, resist exhaustive enumeration. In terms of value, not all creations are valuable. Obtaining something valuable with an exceedingly low probability is impractical, as even a monkey could theoretically print Shakespeare's works. How do LLMs ensure the probability of generating valuable creations? *What are the mechanisms underlying the novelty and value of creation?* A prevalent conjecture posits that **novelty is generated through the random walk** (Sæbø and Brovold, 2024). However, intuitively, **current LLMs themselves seem unable to evaluate the value of creations due to architectural limitations** (Chakrabarty et al., 2024). Because, once the next token is generated, there is no intrinsic mechanism for accepting or rejecting the creations. This hinders the evaluation of the usefulness and value of proposed novelties, as humans do, by bending, blending, or breaking biases (Sæbø and Brovold, 2024). Some works assume that each token is indeed valuable and meets long-term expectations. However, the well-known hallucination problem (Xu et al., 2024d) of LLMs refutes this assumption. Besides, the transformer architecture struggles with long context (Li et al., 2024b), despite the existence of many variants for addressing this issue (Huang et al., 2023c; Liu et al., 2024b). More importantly, MLPs of Transformer may also work contrary to creativity, i.e., the increased attentions narrow the conditional distribution for token prediction (Sæbø and Brovold, 2024).

3 Knowledge Evolution in LLMs

Knowledge in LLMs should evolve with changes in the external environment. Therefore, we introduce the Dynamic Intelligence Hypothesis for knowledge evolution in individuals and groups.

Hypothesis 5: Dynamic Intelligence

Conflict and Integration Coexist in the Dynamic Knowledge Evolution of LLMs.

3.1 Individual Evolution

Immersed in a dynamic world, individuals mature through an iterative process of memorization, forgetting, error correction, and deepening understanding of the world around them. Similarly, LLMs dynamically encapsulate knowledge into parameters through the process of conflict and integration.

In the *pre-training phase*, LLMs start as blank slates, facilitating easier acquisition for new knowledge (Allen-Zhu and Li, 2024; Zhou et al., 2023a). Consequently, numerous experiments demonstrate that LLMs accumulate vast amounts of knowledge during this stage (Cao et al., 2024b; Zhou et al., 2023a; Kaddour et al., 2023; Naveed et al., 2023; Singhal et al., 2022). Later, Akyürek et al. (2022) goes on to identify which training examples are instrumental in endowing LLMs with specific knowledge. However, contradictions during the pre-training stage may induce conflicts among internal parametric knowledge. On the one hand, the false and contradictory information in training corpus propagate and contaminate related memories in LLMs via semantic diffusion, introducing broader detrimental effects beyond direct impacts (Bian et al., 2023). On the other hand, existing LLMs tend to prioritize memorizing more frequent and challenging facts, which can result in subsequent facts overwriting prior memorization, significantly

hindering the memorization of low-frequency facts (Lu et al., 2024). In other words, LLMs struggle with balancing and integrating both low and high-frequency knowledge.

After pre-training, LLMs are anticipated to refresh their internal knowledge to keep pace with the evolving world during *post-training stage*. Although LLMs seem to absorb new knowledge through continued learning, follow user instructions via instruct tuning (Zhang et al., 2023c), and align with human values through alignment tuning (Ziegler et al., 2019), Ji et al. (2024) have noted that LLMs intrinsically resist alignment during the post-training phase. In other words, LLMs tend to learn factual knowledge through pre-training, whereas fine-tuning [3] teaches them to utilize it more efficiently (Gekhman et al., 2024; Zhou et al., 2023a; Ovadia et al., 2024). Ren et al. (2024a) also posit that instruction tuning is a form of self-alignment with existing internal knowledge rather than a process of learning new information. We conjecture that the debate on whether these processes truly introduce new knowledge stems from information conflicts. For example, the conflict between outdated information within LLMs and new external knowledge exacerbates their difficulty in learning new information. To mitigate information conflicts, Ni et al. (2023) propose first forgetting old knowledge then learning new knowledge. Another technique, retrieval-augmented generation (RAG) (Huang and Huang, 2024), while avoiding conflicts within internal parameters, still needs to manage conflicts between retrieved external information and LLMs' internal knowledge (Xu et al., 2024b). RAG also attempt to efficiently and effectively integrate new knowledge across passages or documents using multiple retrieval (Yang et al., 2024a) and hippocampal indexing (Gutiérrez et al., 2024). Besides, editing technologies, including knowledge and representation editing, exhibit promising potential for knowledge addition, modification, and erasure. Specifically, knowledge editing (Meng et al., 2022; Mitchell et al., 2022; Cao et al., 2021b; Zhang et al., 2024a; Wang et al., 2023d; Mazzia et al., 2023) aims to selectively modify model parameters responsible for specific knowledge retention, while representation editing (Zou et al., 2023; Wu et al., 2024) adjusts the model's conceptualization of knowledge to revise the stored knowledge within LLMs. Note that the other strategy for knowledge editing adds external parameters or memory banks for new knowledge while preserving models' parameters. We also provide the comparison of the above methods in §B.2.1 for better understanding.

[3] Fine-tuning includes instruct tuning and alignment tuning (Zhao et al., 2023b).

3.2 Group Evolution

Besides individual learning, social interaction plays a pivotal role in the acquisition of new knowledge and is a key driver of human societal development (Baucal et al., 2014; Levine et al., 1993). LLMs, also known as agents, collaborate to accomplish complex tasks during group evolution, each bearing unique knowledge that may sometimes contradict each other. Therefore, contrary to individual evolution, *group evolution encounters intensified conflicts, such as conflicts in specialized expertise among agents, competing interests, cultural disparities, moral dilemmas, and others*. To achieve consensus and resolve conflicts, agents must first clarify their own and others' goals (beliefs) through internal representations in models (Zhu et al., 2024; Zou et al., 2023). Agents then discuss, debate, and reflect on shared knowledge through various communication methods (Chan et al., 2024; Smit et al., 2024; Li et al., 2024e; Soltoggio et al., 2024), e.g., prompt instructions, task and agent descriptions, parameter signals (activation and gradient), and representations of models. However, conformity of agents, which tends to believe the majority's incorrect answers rather than maintaining their own, hinders conflict resolution during group evolution (Zhang et al., 2023a; Ma et al., 2024). Note that the group also struggles with automating moral decision-making when facing moral conflicts. Specifically, agents in the group miss ground truth for moral "correctness" and encounter dilemmas due to changes in moral norms over time (Hagendorff and Danks, 2023). Generally, when, what, and how to share knowledge in the communication process to maximize learning efficiency and long-term expectations are still open questions in group evolution.

Through debate and collaboration, *groups integrate more knowledge and can surpass the cognition of individual units* (Liang et al., 2023a; Qian et al., 2023; Qiao et al., 2024; Talebirad and Nadiri, 2023; Zhang et al., 2023a). This derives from the assumption that each individual unit can contribute to and benefit from the collective knowledge

(Soltoggio et al., 2024; Xu et al., 2024c). In addition, *"When a measure becomes a target, it ceases to be a good measure"*, which implies that optimizing one objective on a single individual will inevitably harm other optimization objectives to some extent. Hence, it is unrealistic for an individual to learn all knowledge compared to group optimization. Interestingly, LLM groups also follow the collaborative scaling law (Qian et al., 2024a), where normalized solution quality follows a logistic growth pattern as scaling agents. Moreover, some works (Huh et al., 2024; Bereska and Gavves, 2024) propose that knowledge tends to converge into the same representation spaces among the whole artificial neural models group with different data, modalities, and objectives.

Note that the above mechanism analysis of knowledge utilization and evolution may provide an avenue to construct more efficient and trustworthy models in practice. We further elaborate the application and its implications in Appendix §E.

4 Discussion

In this section, we discuss some open questions and seek to explore their essence and underlying principles. Specifically, we discuss what knowledge LLMs have learned in §4.1, examine the fragility of the learned knowledge in application in §4.2, analyze the dark knowledge not yet learned by machines or humans in §4.3, and explore how LLMs can expand the boundaries of unknown knowledge from interdisciplinary perspectives §F.

4.1 What Knowledge Have LLMs Learned?

Critics question whether LLMs truly have knowledge or if they are merely mimicking (Schwarzschild et al., 2024), akin to the "Stochastic Parro" (Bender et al., 2021) and "Clever Hans" (Shapira et al., 2024). We first review the doubts from the following three levels through *observation phenomena*: 1) Memorization: LLMs primarily rely on positional information over semantic understanding (Li et al., 2022) to predict answers. Additionally, LLMs may generate different answers for the same question due to different expressions. 2) Comprehension and application: Allen-Zhu and Li (2023b) argue that LLMs hardly efficiently apply knowledge from pre-training data, even when such knowledge is perfectly stored and fully extracted from LLMs. Therefore, LLMs struggle with various reasoning tasks (Wu et al., 2023b; Nezhurina et al., 2024; Gutiérrez et al., 2024) as well as the reversal curse (Berglund et al., 2023). Besides, LLMs are not yet able to reliably act as text world simulators and encounter difficulties with planning (Wang et al., 2024d). 3) Creation: Although LLMs are capable of generating new terms, their quality often falls below that created by humans (Raiola, 2023). Even though LLMs possess knowledge, some critics argue that current *analysis methods* may only explain low-level co-occurrence patterns, not internal mechanisms. The primary criticism asserts that the components responsible for certain types of knowledge in LLM fail to perform effectively in practical applications (Hase et al., 2023). In addition, the components responsible for specific knowledge within LLMs vary under different methods. For these criticisms, Chen et al. (2024f,d) propose degenerate neurons and posit that different degenerate components indeed independently express a fact. Chen et al. (2024e) delineate the differences in the mechanisms of knowledge storage and representation, proposing the Query Localization Assumption to response these controversies. Zhu and Li (2023) further observe that knowledge may be memorized but not extracted due to the knowledge not being sufficiently augmented (e.g., through paraphrasing, sentence shuffling) during pretraining. Hence, rewriting the training data to provide knowledge augmentation and incorporating more instruction fine-tuning data in the pretraining stage can effectively alleviate the above challenges and criticisms.

Despite considerable criticism, the mainstream view (Didolkar et al., 2024; Jin and Rinard; Jin, 2024) is that **current LLMs may possess basic world knowledge via memorization but hardly master underlying principles for reasoning and creativity**. In other words, LLMs master basic knowledge via memorization (discussed in §2.1). Although LLMs possess the foundational ability to comprehend and apply knowledge (discussed in §2.2), exhibiting plausible and impressive reasoning capabilities. Current LLMs still struggle with reasoning and planning in complex tasks due to the fragility of knowledge in LLMs (elaborated in §4.2). These reasoning and planning abilities usually require to be induced through techniques such as ICL and CoT. Unfortunately, current LLMs are nearly incapable of creation due to the architectural limitations (discussed in §2.3). Therefore, some scholars explore various architectural choices

(e.g., Mamba (Gu and Dao, 2023)) and training procedures. Besides, recent research attempts to manipulate neurons, knowledge circuits, or representations (Allen-Zhu and Li, 2023b; Zou et al., 2023; Wu et al., 2024; Li et al., 2023a) to explore more knowledge and awaken the reasoning and planning capabilities of LLMs.

✨ **Remarks**: LLMs have learned basic knowledge of the world by *momorization*. However, the learned knowledge is fragile, leading to challenges in *knowledge comprehension and application*. Unfortunately, due to architectural limitations, current LLMs struggle with *creation*.

4.2 Why Is Learned Knowledge Fragile?

The knowledge learned by LLMs is fragile, leading to challenges in application including hallucination, knowledge conflicts, failed reasoning, and safety risk [4]. **Hallucination** denotes content generated by LLMs that diverges from real-world facts or inputs (Huang et al., 2023b; Xu et al., 2024d; Farquhar et al., 2024; Chen et al., 2024c). On the one hand, factuality hallucination underscores the disparity between generated content and real-world knowledge. On the other hand, faithfulness hallucination describes the departure of generated content from user instructions or input context, as well as the coherence maintained within the generated content. **Knowledge Conflict** inherently denotes inconsistencies in knowledge (Xu et al., 2024b; Kortukov et al., 2024). On the one hand, internal memory conflicts within the model cause LLMs to exhibit unpredictable behaviors and generate differing results to inputs which are semantically equivalent but syntactically distinct (Xu et al., 2024b; Wang et al., 2023a; Feng et al., 2023b; Raj et al., 2022). On the other hand, context-memory conflict emerges when external context knowledge contradicts internal parametric knowledge (Xu et al., 2024b; Mallen et al., 2023).

We posit that **these challenges mainly derive from improper learning data.** Specifically, hallucination is introduced by data (Kang and Choi, 2023; Weng, 2024; Zhang et al., 2024c), heightened during the pre-training (Brown et al., 2020; Chiang and Cholak, 2022), alignment (Azaria and Mitchell, 2023; Ouyang et al., 2022), and deficiencies in decoding strategies (Fan et al., 2018; Chuang et al., 2023; Shi et al., 2023). Internal memory conflict can be attributed to training corpus bias (Wang et al., 2023b), and exacerbated by decoding strategies (Lee et al., 2022b) and knowledge editing. Context-memory conflict arises mainly from the absence of accurate knowledge during training, necessitating retrieval from databases and the Web. Failed reasoning usually arises from improper **data distribution**. Specifically, knowledge may be memorized but not extractable or applicable without sufficient augmentation (e.g., through paraphrasing, sentence shuffling) during pre-training (Zhu and Li, 2023). Antoniades et al. (2024) also delve into the mechanism between parametric knowledge and learning data, demonstrate that training data distribution qualitatively influences generalization behavior (Jiang et al., 2024a). Wang et al. (2024a) further suggest that improper data distribution in the corpus causes LLMs to lack essential reasoning components, such as the bridge layer for two-hop reasoning. Similar mechanism analysis also supports the above conclusion, indicating that hallucinations arise from a lack of mover heads (Yao et al., 2024; Yu et al., 2024b), while knowledge conflicts stem from circuit competition failure in the last few layers (Lv et al., 2024; Merullo et al., 2023b; Hase et al., 2023; Ju et al., 2024; Jin et al., 2024b). Additionally, **data quantity** is crucial for knowledge robustness. Specifically, LLMs can systematically learn comprehensive understandings of the world from extensive datasets, while little data during post-training stage may compromise the robustness of knowledge representation. This assumption is confirmed by numerous failures of post-training. For example, SFT exacerbates hallucinations (Gekhman et al., 2024; Kang et al., 2024), and knowledge editing amplifies knowledge conflicts (Li et al., 2023d; Yang et al., 2024c). Note that safety issues usually caused by the distribution of unseen data (adversarial input) (Wei et al., 2023; Li et al., 2024c), which is elaborated in §E.2.

✨ **Remarks**: Improper learning caused by data distribution and quantity might be the fundamental and primary cause.

4.3 Does Difficult-to-Learn "Dark Knowledge" Exist?

The distribution and quality of data are vital for knowledge acquisition and robust operation within the model (machine). Imagine an ideal scenario where we have access to all kinds of data to train the machine. The data includes all possible modali-

[4] The secure risk is elaborated in §E.2.

ties, such as text, image, audio, video, etc. Models can also interact with each other and the external environment. In this long-term development, will there still be unknown dark knowledge for intelligence to human or model (machine)?

We hypothesize that there will still **exist dark knowledge for intelligence** in the future. As shown in Fig 3, dark knowledge describes knowledge unknown to human or machine from the following three situations: 1) knowledge unknown to human & known to machine (UH, KM). Machines leverage vast amounts of data to explore internal patterns, whereas humans struggle with processing such data due to physiological limitations on data processing capacity and computational limits (Burns et al., 2023; McAleese et al., 2024). (UH, KM) includes gene prediction, intelligent transportation systems, and more. Specifically, the structural elucidation of proteins remains mysterious to humans for a long time. Cryo-electron microscopy, through capturing millions of images, first reveals the three-dimensional structures of proteins. Now, neural models can directly predict protein properties with high efficiency and accuracy (Pak et al., 2023). 2) knowledge known to human & unknown to machine (UH, KM). On the one hand, some scholars claim that machine can possess a "Theory of Mind" capability (Zhu et al., 2024) and emotions (Normoyle et al., 2024). On the other hand, critics contend that machine lacks sentience (Alvero and Peña, 2023) and merely probabilistically generates tokens. The causes, extent, and dynamics of these emotions and sentience (like hunger, happiness, and loneliness) are subtle and intricate, making precise mathematical modeling by the machine exceptionally challenging. Specifically, different factors are tightly coupled, making it nearly impossible to disentangle clear input-output relationships as with well-defined factual knowledge. The sentient knowledge also exhibits chaotic behavior (Li et al., 2020; Debbouche et al., 2021), being highly sensitive to initial conditions, where small changes can lead to vastly different outcomes (Segretain et al., 2020). Therefore, opponents argue that no matter how many parameters machine possesses, it cannot learn all the knowledge that human has mastered. 3) knowledge unknown to human & unknown to machine (UH, UM) is beyond our cognition, e.g., the uncertainty in quantum mechanics and the origin of the universe. Generally, Dark knowledge extends beyond current data and model

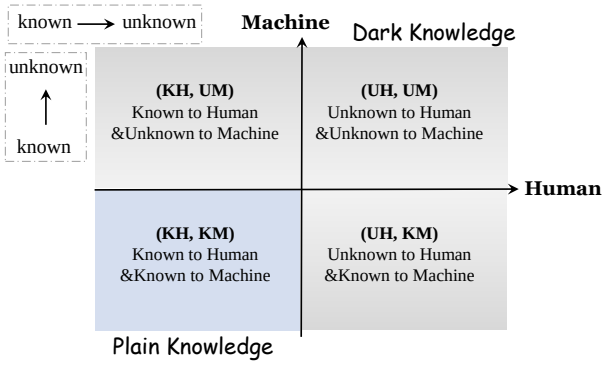

Figure 3: The future cognition of knowledge. The direction of the arrow represents the transition of knowledge from known to unknown. Dark knowledge, represented in gray, denotes knowledge unknown to human or machine. Plain knowledge known to both human and machine is highlighted in blue.

architectures (Tseng et al., 2024). (UH, UM) necessitates human-machine collaboration. Yet, there is no definitive conclusion on whether (UH, KM) and (KH, UM) will be solved by model architecture, training data, and computational resources. Note that plain knowledge known to human and machine in Fig 3 encompasses well-defined historical events, mathematical theorems, physical laws, etc.

✨ **Remarks**: Dark knowledge may persist for a long time and requires human-machine collaboration to explore.

Note that we also discuss some avenues to narrow the boundaries of dark knowledge from interdisciplinary insights in the Appendix §F

5 Conclusion

In this paper, we propose a novel knowledge mechanism analysis taxonomy and review knowledge evolution. We further explore the knowledge LLMs have learned, assess its fragility, and analyze unknown dark knowledge. As for future works, we discuss some promising directions, including Parametric VS. Non-Parametric Knowledge (§G.1), Embodied Intelligence (§G.2), and Domain LLMs (§G.3). We hope these insights may inspire some promising directions for future research and shed light on more powerful and trustworthy models.

Limitations

This work has some limitations as follows:

Hypothesis Despite reviewing a large body of literature and proposing several promising hypotheses, there are still some limitations. On the one

hand, there may be other hypotheses for knowledge utilization and evolution in LLMs. On the other hand, the accuracy of these hypotheses requires further exploration and validation over time.

Knowledge There are various forms of knowledge representation. However, due to current research constraints, this paper does not delve into space (Li et al., 2024f), time (Gurnee and Tegmark, 2023), event-based knowledge, and geoscience (Lin et al., 2024).

Reference The field of knowledge mechanisms is developing rapidly and this paper may miss some important references. Additionally, due to the page limit, we have omit certain technical details. We will continue to pay attention to and supplement new works.

Models Despite mentioning artificial neural models in this paper, knowledge mechanism analysis focuses on LLMs. We will continue to pay attention to other modal models progresses. Besides, all existing work has not considered models larger than 100 billion parameters. Whether the knowledge mechanisms within large-scale models are consistent with smaller ones remains to be studied.

Ethics Statement

We anticipate no ethical or societal implications arising from our research. However, we acknowledge that the internal mechanisms of large language models might be exploited for malicious purposes. We believe such malicious applications can be prevented through model access and legislative regulation. More critically, a transparent model contributes to the development of safer and more reliable general artificial intelligence.

Acknowledgements

We would like to express gratitude to the anonymous reviewers for their kind comments. This work was supported by the National Natural Science Foundation of China (No. 62206246, No. NSFCU23B2055, No. NSFCU19B2027), the Fundamental Research Funds for the Central Universities (226-2023-00138), Zhejiang Provincial Natural Science Foundation of China (No. LGG22F030011), Yongjiang Talent Introduction Programme (2021A-156-G), Information Technology Center and State Key Lab of CAD&CG, Zhejiang University, and NUS-NCS Joint Laboratory (A-0008542-00-00).

References

Mohammad Abufadda and Khalid Mansour. 2021. A survey of synthetic data generation for machine learning. In *22nd International Arab Conference on Information Technology, ACIT 2021, Muscat, Oman, December 21-23, 2021*, pages 1–7. IEEE.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4895–4901. Association for Computational Linguistics.

Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. 2024. Evolutionary optimization of model merging recipes. *CoRR*, abs/2403.13187.

Ekin Akyürek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards tracing factual knowledge in language models back to the training data. *arXiv preprint arXiv:2205.11482*.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023a. Physics of language models: Part 1, context-free grammar. *CoRR*, abs/2305.13673.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023b. Physics of language models: Part 3.2, knowledge manipulation. *CoRR*, abs/2309.14402.

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.3, knowledge capacity scaling laws. *CoRR*, abs/2404.05405.

Aj Alvero and Courtney Peña. 2023. AI sentience and socioculture. *J. Soc. Comput.*, 4(3):205–220.

Antonis Antoniades, Xinyi Wang, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 2024. Generalization v.s. memorization: Tracing language models' capabilities back to pretraining data. *arXiv preprint arXiv:2407.14985*.

Marianna Apidianaki. 2023. From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. *Comput. Linguistics*, 49(2):465–523.

Amos Azaria and Tom M. Mitchell. 2023. The internal state of an LLM knows when its lying. *CoRR*, abs/2304.13734.

Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. 2022. Molgpt: Molecular generation using a transformer-decoder model. *J. Chem. Inf. Model.*, 62(9):2064–2076.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan,

Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Andrea Baronchelli, Ramon Ferrer-i Cancho, Romualdo Pastor-Satorras, Nick Chater, and Morten H Christiansen. 2013. Networks in cognitive science. *Trends in cognitive sciences*, 17(7):348–360.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3319–3327. IEEE Computer Society.

Aleksandar Baucal, Lausanne Mouline, Switzerland Kristiina Kumpulainen, Charis Psaltis, and Baruch Schwarz. 2014. Social interaction in learning and development.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Comput. Linguistics*, 48(1):207–219.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *CoRR*, abs/2303.08112.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yoshua Bengio. 2024. Government interventions to avert future catastrophic ai risks. *Harvard Data Science Review*, (Special Issue 5).

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. 2024. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845.

Yoshua Bengio and Nikolay Malkin. 2024. Machine learning and information theory concepts towards an AI mathematician. *CoRR*, abs/2403.04571.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for AI safety - A review. *CoRR*, abs/2404.14082.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *CoRR*, abs/2309.12288.

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. 2024. Oceangpt: A large language model for ocean science tasks.

Ning Bian, Peilin Liu, Xianpei Han, Hongyu Lin, Yaojie Lu, Ben He1, and Le Sun. 2023. A drop of ink may make a million think: The spread of false information in large language models. *arxiv.org/pdf/2305.04812v1*.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Alberto Blanco-Justicia, Najeeb Jebreel, Benet Manzanares-Salor, David Sánchez, Josep Domingo-Ferrer, Guillem Collell, and Kuan Eeik Tan. 2024. Digital forgetting in large language models: A survey of unlearning methods. *CoRR*, abs/2404.02062.

Benjamin S. Bloom, Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. David McKay Company.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Tristan Hume Josiah E. Burke, Shan Carter, Tom Henighan, , and Chris Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

John N. A. Brown and Lukas Esterle. 2020. I'm already optimal: the dunning-kruger effect, sociogenesis, and self-integration. In *2020 IEEE International Conference on Autonomic Computing and Self-Organizing Systems, ACSOS 2020, Companion Volume, Washington, DC, USA, August 17-21, 2020*, pages 82–84. IEEE.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *CoRR*, abs/2312.09390.

Tolga Buz, Benjamin Frost, Nikola Genchev, Moritz Schneider, Lucie-Aimée Kaffee, and Gerard de Melo. 2024. Investigating wit, creativity, and detectability of large language models in domain-specific writing style adaptation of reddit's showerthoughts. *CoRR*, abs/2405.01660.

Nitay Calderon and Roi Reichart. 2024. On behalf of the stakeholders: Trends in nlp model interpretability in the era of llms. *arXiv preprint arXiv:2407.19200*.

Boxi Cao, Hongyu Lin, Xianpei Han, and Le Sun. 2024a. The life cycle of knowledge in big language models: A survey. *Mach. Intell. Res.*, 21(2):217–238.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021a. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1860–1874. Association for Computational Linguistics.

Boxi Cao, Qiaoyu Tang, Hongyu Lin, Shanshan Jiang, Bin Dong, Xianpei Han, Jiawei Chen, Tianshu Wang, and Le Sun. 2024b. Retentive or forgetful? diving into the knowledge memorizing mechanism of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14016–14036, Torino, Italia. ELRA and ICCL.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021b. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 30:1–30:34. ACM.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Lawrence Chan, Leon Lang, and Erik Jenner. 2023. Natural abstractions: Key claims, theorems, and critiques.

Huajun Chen. 2024. Large knowledge model: Perspectives and challenges. *Data Intelligence*.

Lihu Chen, Adam Dejl, and Francesca Toni. 2024a. Analyzing key neurons in large language models. *arXiv preprint arXiv:2406.10868*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024b. Are we on the right way for evaluating large vision-language models? *CoRR*, abs/2403.20330.

Xiaohan Chen, Yu Cheng, Shuohang Wang, Zhe Gan, Jingjing Liu, and Zhangyang Wang. 2021. The elastic lottery ticket hypothesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 26609–26621.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David F Fouhey, and Joyce Chai. 2024c. Multi-object hallucination in vision-language models. *arXiv preprint arXiv:2407.06192*.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024d. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17817–17825. AAAI Press.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024e. Knowledge localization: Mission not accomplished? enter query localization! *CoRR*, abs/22405.14117.

Yuheng Chen, Pengfei Cao, Yubo Chen, Yining Wang, Shengping Liu, Kang Liu, and Jun Zhao. 2024f. The da vinci code of large pre-trained language models: Deciphering degenerate knowledge neurons. *CoRR*, abs/2402.13731.

David Chiang and Peter Cholak. 2022. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7654–7664. Association for Computational Linguistics.

Yejin Choi. 2022. Knowledge is power: Symbolic knowledge distillation, commonsense morality, & multimodal script knowledge. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, page 3. ACM.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *CoRR*, abs/2309.03883.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. 2023. A toy model of universality: Reverse engineering how networks learn group operations. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 6243–6267. PMLR.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models. *CoRR*, abs/2307.12976.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Carlos Coronel-Oliveros, Vicente Medel, Sebastián Orellana, Julio Rodiño, Fernando Lehue, Josephine Cruzat, Enzo Tagliazucchi, Aneta Brzezicka, Patricio Orio, Natalia Kowalczyk-Grebska, and Agustín Ibáñez. 2024. Gaming expertise induces meso-scale brain plasticity and efficiency mechanisms as revealed by whole-brain modeling. *NeuroImage*, 293:120633.

Joy Crosbie and Ekaterina Shutova. 2024. Induction heads as an essential mechanism for pattern matching in in-context learning. *arxiv.org/pdf/2407.07011*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *CoRR*, abs/2309.08600.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark W. Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua B. Tenenbaum. 2024. Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems. *CoRR*, abs/2405.06624.

Fahim Dalvi, Hassan Sajjad, and Nadir Durrani. 2023. Neurox library for neuron analysis of deep NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023, Toronto, Canada, July 10-12, 2023*, pages 226–234. Association for Computational Linguistics.

Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, Aimin Zhou, Ze Zhou, Qin Chen, Jie Zhou, Liang He, and Xipeng Qiu. 2023. Educhat: A large-scale language model-based chatbot system for intelligent education. *CoRR*, abs/2308.02773.

Randall Davis, Howard E. Shrobe, and Peter Szolovits. 1993. What is a knowledge representation? *AI Mag.*, 14(1):17–33.

Thiebaut de Schotten, Michel, Forkel, and Stephanie J. 2022. The emergent properties of the connected brain. *Science*, 378(6619):505–510.

Nadjette Debbouche, Adel Ouannas, Iqbal M. Batiha, Giuseppe Grassi, Mohammed K. A. Kaabar, Hadi Jahanshahi, Ayman A. Aly, and Awad M. Aljuaid. 2021. Chaotic behavior analysis of a new incommensurate fractional-order hopfield neural network system. *Complex.*, 2021:3394666:1–3394666:11.

Matthew DeLorenzo, Vasudev Gohil, and Jeyavijayan Rajendran. 2024. Creativeval: Evaluating creativity of llm-based hardware code generation. *CoRR*, abs/2404.08806.

Cheng Deng, Tianhang Zhang, Zhongmou He, Yi Xu, Qiyuan Chen, Yuanyuan Shi, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, Zhouhan Lin, and Junxian He. 2023. K2: A foundation language model for geoscience knowledge understanding and utilization.

Nolan Dey, Gurpreet Gosal, Zhiming Chen, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. 2023. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster. *CoRR*, abs/2304.03208.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy P. Lillicrap, Danilo J. Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *CoRR*, abs/2405.12205.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. 2024. Jump to conclusions: Short-cutting transformers with linear transformations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 9615–9625. ELRA and ICCL.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Raymond Douglas, Andis Draguns, and Tomas Gavenciak. 2024. Mitigating the problem of strong priors in lms with context extrapolation. *CoRR*, abs/2401.17692.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. Transcoders find interpretable llm feature circuits. *CoRR*.

Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. 2024. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *CoRR*, abs/2402.18312.

Carl Edwards, Tuan Manh Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 375–413. Association for Computational Linguistics.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Trans. Assoc. Comput. Linguistics*, 9:160–175.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *CoRR*, abs/2209.10652.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: model alignment as prospect theoretic optimization. *CoRR*, abs/2402.01306.

Jose A Fadul. 2009. Collective learning: Applying distributed cognition for collective intelligence. *International Journal of Learning*, 16(4).

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics.

Yin Fang, Ningyu Zhang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. Domain-agnostic molecular generation with self-feedback. *CoRR*, abs/2301.11259.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023a. Aligning semantic in brain and language: A curriculum contrastive method for electroencephalography-to-text generation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Zhangyin Feng, Weitao Ma, Weijiang Yu, Lei Huang, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023b. Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications. *CoRR*, abs/2311.05876.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models. *CoRR*, abs/2405.00208.

Sharlene N Flesher, John E Downey, Jeffrey M Weiss, Christopher L Hughes, Angelica J Herrera, Elizabeth C Tyler-Kabara, Michael L Boninger, Jennifer L Collinger, and Robert A Gaunt. 2021. A brain-computer interface that evokes tactile sensations improves robotic arm control. *Science*, 372(6544):831–836.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024a. Scaling and evaluating sparse autoencoders. *CoRR*, abs/2406.04093.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024b. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Leo Gao, Tom Dupré la Tour andHenk Tillman, Gabriel Goh andRajan Troll, Alec Radford andIlya Sutskever, Jan Leike, and Jeffrey Wu. 2024c. Scaling and evaluating sparse autoencoders. *CoRR*, abs/2406.04093.

Huaizhi Ge, Frank Rudzicz, and Zining Zhu. 2024. What do the circuits mean? a knowledge edit view. *arXiv preprint arXiv:2406.17241*.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. 2022. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *CoRR*, abs/2405.05904.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12216–12235. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir, and Yoav Goldberg. 2022a. Lm-debugger: An interactive tool for inspection and intervention in transformer-based language models. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 - System Demonstrations, Abu Dhabi, UAE, December 7-11, 2022*, pages 12–21. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022b. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 30–45. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's mergekit: A toolkit for merging large language models. *CoRR*, abs/2403.13257.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching. *CoRR*, abs/2304.05969.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of llms on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 14504–14528. Association for Computational Linguistics.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680.

C. A. E. Goodhart. 1984. *Problems of Monetary Management: The UK Experience*, pages 91–121. Macmillan Education UK, London.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. 2023a. Successor heads: Recurring, interpretable attention heads in the wild. *CoRR*, abs/2312.09230.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. 2023b. Successor heads: Recurring, interpretable attention heads in the wild. *CoRR*, abs/2312.09230.

Colin M. Gray, Cristiana Teixeira Santos, Nataliia Bielova, and Thomas Mildner. 2024. An ontology of dark patterns knowledge: Foundations, definitions, and a pathway for shared knowledge-building. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 289:1–289:22. ACM.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *CoRR*, abs/2312.00752.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model editing can hurt general abilities of large language models. *CoRR*, abs/2401.04700.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip H. S. Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. *CoRR*, abs/2307.12980.

Anshita Gupta, Debanjan Mondal, Akshay Krishna Sheshadri, Wenlong Zhao, Xiang Li, Sarah Wiegreffe, and Niket Tandon. 2023. Editing common sense in transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 8214–8232. Association for Computational Linguistics.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *CoRR*, abs/2305.01610.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *CoRR*, abs/2310.02207.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *CoRR*, abs/2405.14831.

Jennifer Haase and Paul H. P. Hanel. 2023. Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *CoRR*, abs/2303.12003.

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *CoRR*, abs/2303.13988.

Thilo Hagendorff and David Danks. 2023. Ethical and methodological challenges in building morally informed AI systems. *AI Ethics*, 3(2):553–566.

Xu Han, Zhengyan Zhang, and Zhiyuan Liu. 2021. Knowledgeable machine learning for natural language processing. *Commun. ACM*, 64(11):50–51.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2022. PTR: prompt tuning with rules for text classification. *AI Open*, 3:182–192.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Peter Hase, Thomas Hofweber, Xiang Zhou, Elias Stengel-Eskin1, and Mohit Bansal1. 2024. Fundamental problems with model editing: How should rational belief revision work in llms? *arxiv.org/pdf/2406.19354*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Qiyuan He, Yizhong Wang, and Wenya Wang. 2024a. Can language models act as knowledge bases at scale? *CoRR*, abs/2402.14273.

Zhengfu He, Xuyang Ge, Qiong Tang, Tianxiang Sun, Qinyuan Cheng, and Xipeng Qiu. 2024b. Dictionary learning improves patch-free circuit discovery in mechanistic interpretability: A case study on othello-gpt. *CoRR*, abs/2402.12201.

Kenneth M Heilman, Stephen E Nadeau, and David O Beversdorf. 2003. Creative innovation: possible brain mechanisms. *Neurocase*, 9(5):369–379.

Benjamin Heinzerling and Kentaro Inui. 2020. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. *CoRR*, abs/2008.09036.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exbert: A visual analysis tool to explore learned representations in transformer models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 187–196. Association for Computational Linguistics.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the*

2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pages 4902–4919. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yi Hu, Xiaojuan Tang, Haotong Yang, and Muhan Zhang. 2024. Case-based or rule-based: How do transformers do the math? *CoRR*, abs/2402.17709.

Han Huang, Haitian Zhong, Tao Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2024a. Vlkeb: A large vision-language model knowledge editing benchmark.

Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. 2023a. Rigorously assessing natural language explanations of neurons. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2023, Singapore, December 7, 2023*, pages 317–331. Association for Computational Linguistics.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. 2024b. RAVEL: evaluating interpretability methods on disentangling language model representations. *CoRR*, abs/2402.17700.

Jing Huang, Diyi Yang, and Christopher Potts. 2024c. Demystifying verbatim memorization in large language models. *CoRR*, abs/2407.17817.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232.

Yizheng Huang and Jimmy Huang. 2024. A survey on retrieval-augmented text generation for large language models. *CoRR*, abs/2404.10981.

Yunpeng Huang, Jingwei Xu, Zixu Jiang, Junyu Lai, Zenan Li, Yuan Yao, Taolue Chen, Lijuan Yang, Zhou Xin, and Xiaoxing Ma. 2023c. Advancing transformer architecture in long-context large language models: A comprehensive survey. *CoRR*, abs/2311.12351.

Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. 2024d. Compression represents intelligence linearly. *CoRR*, abs/2404.09937.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

John Hyman. 1999. How knowledge works. *The philosophical quarterly*, 49(197):433–451.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Mohsen Jamali, Benjamin Grannan, Jing Cai, Arjun R Khanna, William Muñoz, Irene Caprara, Angelique C Paulk, Sydney S Cash, Evelina Fedorenko, and Ziv M Williams. 2024. Semantic encoding during language comprehension at single-cell resolution. *Nature*, pages 1–7.

Adam S. Jermyn, Nicholas Schiefer, and Evan Hubinger. 2022. c. *CoRR*, abs/2211.09169.

Jiaming Ji, Kaile Wang, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, and Yaodong Yang. 2024. Language models resist alignment. *arXiv preprint arXiv:2406.06144*.

Shuyang Jiang, Yusheng Liao, Ya Zhang, Yu Wang, and Yanfeng Wang. 2024a. Taia: Large language models are out-of-distribution data learners. *arXiv preprint arXiv:2405.20192*.

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. 2024b. Do llms dream of elephants (when told not to)? latent concept association and associative memory in transformers. *CoRR*.

Licheng Jiao, Zhongjian Huang, Xu Liu, Yuting Yang, Mengru Ma, Jiaxuan Zhao, Chao You, Biao Hou, Shuyuan Yang, Fang Liu, Wenping Ma, Lingling Li, Puhua Chen, Zhixi Feng, Xu Tang, Yuwei Guo, Xiangrong Zhang, Dou Quan, Shuang Wang, Weibin Li, Jing Bai, Yangyang Li, Ronghua Shang, and Jie Feng. 2023. Brain-inspired remote sensing interpretation: A comprehensive survey. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 16:2992–3033.

Charles Jin. 2024. Latent causal probing: A formal perspective on probing with causal models of data. *arXiv preprint arXiv:2407.13765*.

Charles Jin and Martin Rinard. Emergent representations of program semantics in language models trained on programs. In *Forty-first International Conference on Machine Learning*.

Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. 2024a. Exploring concept depth: How large language models acquire knowledge at different layers? *CoRR*, abs/2404.07066.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024b. Cutting off the head ends the

conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. *CoRR*, abs/2402.18154.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? A layer-wise probing study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 8235–8246. ELRA and ICCL.

Jean Kaddour. 2023. The minipile challenge for data-efficient language models. *CoRR*, abs/2304.08442.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *CoRR*, abs/2307.10169.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Cheongwoong Kang and Jaesik Choi. 2023. Impact of co-occurrence on factual knowledge of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7721–7735. Association for Computational Linguistics.

Katie Kang, Eric Wallace, Claire J. Tomlin, Aviral Kumar, and Sergey Levine. 2024. Unfamiliar finetuning examples control how language models hallucinate. *CoRR*, abs/2403.05612.

Adam Karvonen, Benjamin Wright, Can Rager, Rico Angell, Jannik Brinkmann, Logan Smith, Claudio Mayrink Verdun, David Bau, and Samuel Marks. 2024. Measuring progress in dictionary learning for language model interpretability with board game models. *arXiv preprint arXiv:2408.00113*.

Mehran Kazemi, Sid Mittal, and Deepak Ramachandran. 2023. Understanding finetuning for factual knowledge extraction from language models. *CoRR*, abs/2301.11293.

Judith Keene, John Colvin, and Justine Sissons. 2010. Mapping student information literacy activity against bloom's taxonomy of cognitive skills. *Journal of information literacy*, 4(1):6–20.

Miyoung Ko, Sue Hyun Park, Joonsuk Park, and Minjoon Seo. 2024. Investigating how large language models leverage internal knowledge to perform complex reasoning. *arXiv preprint arXiv:2406.19502*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2023. Analyzing feed-forward blocks in transformers through the lens of attention map. *arXiv preprint arXiv:2302.00456*.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR.

Daniel Kolak, William Hirstein, Peter Mandik, and Jonathan Waskan. 2006. *Cognitive science: An introduction to mind and brain*. Routledge.

Arinbjorn Kolbeinsson, Kyle O'Brien, Tianjin Huang, Shanghua Gao, Shiwei Liu, Jonathan Richard Schwarz, Anurag Vaidya, Faisal Mahmood, Marinka Zitnik, Tianlong Chen, et al. 2024. Composable interventions for language models. *arXiv preprint arXiv:2407.06483*.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, Yong Cheng, Ming-Chang Chiu, Josh Dillon, Irfan Essa, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, David A. Ross, Grant Schindler, Mikhail Sirotenko, Kihyuk Sohn, Krishna Somandepalli, Huisheng Wang, Jimmy Yan, Ming-Hsuan Yang, Xuan Yang, Bryan Seybold, and Lu Jiang. 2023. Videopoet: A large language model for zero-shot video generation. *CoRR*, abs/2312.14125.

Lingkai Kong, Haorui Wang, Wenhao Mu, Yuanqi Du, Yuchen Zhuang, Yifei Zhou, Yue Song, Rongzhi Zhang, Kai Wang, and Chao Zhang. 2024. Aligning large language models with representation editing: A control perspective. *arXiv preprint arXiv:2406.05954*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. Studying large language model behaviors under realistic knowledge conflicts. *CoRR*, abs/2404.16032.

Tanya Kraljic and Michal Lahav. 2024. From prompt engineering to collaborating: A human-centered approach to AI interfaces. *Interactions*, 31(3):30–35.

Anastasia Kritharoula, Maria Lymperaiou, and Giorgos Stamou. 2023. Language models as knowledge bases for visual word sense disambiguation. In *Joint proceedings of the 1st workshop on Knowledge Base Construction from Pre-Trained Language Models (KBC-LM) and the 2nd challenge on Language Models for Knowledge Base Construction (LM-KBC) co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November*

6, 2023, volume 3577 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2023. Large language models in law: A survey.

Michael Lan, Fazl, and Barez. 2024. Interpreting shared circuits for ordered sequence prediction in a large language model. *arXiv preprint arXiv:2311.04131*.

Anna Langedijk, Hosein Mohebbi, Gabriele Sarti, Willem H. Zuidema, and Jaap Jumelet. 2023. Decoderlens: Layerwise interpretation of encoder-decoder transformers. *CoRR*, abs/2310.03686.

LawrenceC, Adrià Garriga-alonso, Nicholas Goldowsky Dill, ryan greenblatt, jenny, Ansh Radhakrishnan, Buck, and Nate Thomas. 2022. Causal scrubbing: A method for rigorously testing interpretability hypotheses.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. 2024. A mechanistic understanding of alignment algorithms: A case study on DPO and toxicity. *CoRR*, abs/2401.01967.

Dong-Ho Lee, Akshen Kadakia, Brihi Joshi, Aaron Chan, Ziyi Liu, Kiran Narahari, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, and Xiang Ren. 2023a. XMD: an end-to-end framework for interactive explanation-based debugging of NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023, Toronto, Canada, July 10-12, 2023*, pages 264–273. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023b. RLAIF: scaling reinforcement learning from human feedback with AI feedback. *CoRR*, abs/2309.00267.

Honglak Lee, Alexis J. Battle, Rajat Raina, and Andrew Y. Ng. 2006. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 801–808. MIT Press.

Jin Hyung Lee, Qin Liu, and Ehsan Dadgar-Kiani. 2022a. Solving brain circuit function and dysfunction with computational modeling and optogenetic fmri. *Science*, 378(6619):493–499.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022b. Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Florin Leon. 2024. A review of findings from neuroscience and cognitive psychology as possible inspiration for the path to artificial general intelligence. *CoRR*, abs/2401.10904.

Michael A. Lepori, Ellie Pavlick, and Thomas Serre. 2023. Neurosurgeon: A toolkit for subnetwork analysis. *CoRR*, abs/2309.00244.

John M Levine, Lauren B Resnick, and E Tory Higgins. 1993. Social foundations of cognition. *Annual review of psychology*, 44(1):585–612.

Aaron J Li, Satyapriya Krishna, and Himabindu Lakkaraju. 2024a. More rlhf, more trust? on the impact of human preference alignment on language model trustworthiness. *arXiv preprint arXiv:2404.18870*.

Guohui Li, Xiangyu Zhang, and Hong Yang. 2020. Complexity analysis and synchronization control of fractional-order jafari-sprott chaotic system. *IEEE Access*, 8:53360–53373.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Maximilian Li, Xander Davies, and Max Nadeau. 2023b. Circuit breaking: Removing model behaviors with targeted ablation. *CoRR*, abs/2309.05973.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Chengjie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022. How pre-trained language models capture factual knowledge? A causal-inspired analysis. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1720–1732. Association for Computational Linguistics.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024b. Long-context llms struggle with long in-context learning. *CoRR*, abs/2404.02060.

Tianlong Li, Xiaoqing Zheng, and Xuanjing Huang. 2024c. Open the pandora's box of llms: Jailbreaking llms through representation engineering. *CoRR*, abs/2401.06824.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2024d. PMET: precise model editing in a transformer. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18564–18572. AAAI Press.

Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. 2023c. Diffusion models for image restoration and enhancement - A comprehensive survey. *CoRR*, abs/2308.09388.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024e. Improving multi-agent debate with sparse communication topology. *CoRR*, abs/2406.11776.

Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. 2024f. Urbangpt: Spatio-temporal large language models. *CoRR*, abs/2403.00813.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2023d. Unveiling the pitfalls of knowledge editing for large language models. *CoRR*, abs/2310.02129.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023a. Encouraging divergent thinking in large language models through multi-agent debate. *CoRR*, abs/2305.19118.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023b. Encouraging divergent thinking in large language models through multi-agent debate. *CoRR*, abs/2305.19118.

Zhouhan Lin, Cheng Deng, Le Zhou, Tianhang Zhang, Yi Xu, Yutong Xu, Zhongmou He, Yuanyuan Shi, Beiya Dai, Yunchong Song, Boyi Zeng, Qiyuan Chen, Tao Shi, Tianyu Huang, Yiwei Xu, Shu Wang, Luoyi Fu, Weinan Zhang, Junxian He, Chao Ma, Yunqiang Zhu, Xinbing Wang, and Chenghu Zhou. 2024. Geogalactica: A scientific large language model in geoscience. *CoRR*, abs/2401.00434.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. 2023a. Sophia: A scalable stochastic second-order optimizer for language model pretraining. *CoRR*, abs/2305.14342.

Huanshuo Liu, Hao Zhang, Zhijiang Guo, Kuicai Dong, Xiangyang Li, Yi Quan Lee, Cong Zhang, and Yong Liu. 2024a. Ctrla: Adaptive retrieval-augmented generation via probe-guided control. *CoRR*, abs/2405.18727.

Xiaoran Liu, Qipeng Guo, Yuerong Song, Zhigeng Liu, Kai Lv, Hang Yan, Linlin Li, Qun Liu, and Xipeng Qiu. 2024b. Farewell to length extrapolation, a training-free infinite context with finite attention scope. *arXiv preprint arXiv:2407.15176*.

Ziming Liu, Eric Gan, and Max Tegmark. 2024c. Seeing is believing: Brain-inspired modular training for mechanistic interpretability. *Entropy*, 26(1):41.

Ziming Liu, Mikail Khona, Ila R. Fiete, and Max Tegmark. 2023b. Growing brains: Co-emergence of anatomical and functional modularity in recurrent neural networks. *CoRR*, abs/2310.07711.

Ziyao Liu, Huanyi Ye, Chen Chen, and Kwok-Yan Lam. 2024d. Threats, attacks, and defenses in machine unlearning: A survey. *CoRR*, abs/2403.13682.

Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuanjing Huang, and Xipeng Qiu. 2024. Scaling laws for fact memorization of large language models. *arXiv preprint arXiv:2406.15720*.

Daniel Lundström, Tianjian Huang, and Meisam Razaviyayn. 2022. A rigorous study of integrated gradients method and extensions to internal neuron attributions. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 14485–14508. PMLR.

Haoyan Luo and Lucia Specia. 2024. From understanding to utilization: A survey on explainability for large language models. *CoRR*, abs/2401.12874.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023. Reasoning on graphs: Faithful and interpretable large language model reasoning. *CoRR*, abs/2310.01061.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. 2024a. In-context learning with retrieved demonstrations for language models: A survey. *CoRR*, abs/2401.11624.

Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. 2024b. Local interpretations for explainable natural language processing: A survey. *ACM Comput. Surv.*, 56(9):232:1–232:36.

Ang Lv, Kaiyi Zhang, Yuhan Chen, Yulong Wang, Lifeng Liu, Ji-Rong Wen, Jian Xie, and Rui Yan. 2024. Interpreting key mechanisms of factual recall in transformer-based language models. *CoRR*, abs/2403.19521.

Chenglong Ma, Yongli Ren, Pablo Castells, and Mark Sanderson. 2024. Temporal conformity-aware hawkes graph network for recommendations. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 3185–3194. ACM.

Ali Mahmoodi, Majid Nili Ahmadabadi, and Bahador Bahrami. 2013. The less you know, you think you know more; dunning and kruger effect in collective decision making. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society, CogSci 2013, Berlin, Germany, July 31 - August 3, 2013*. cognitivesciencesociety.org.

Kyle Mahowald, Anna A. Ivanova, Idan Asher Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *CoRR*, abs/2301.06627.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics.

Arthur B Markman. 2013. *Knowledge representation*. Psychology Press.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *CoRR*, abs/2403.19647.

Vittorio Mazzia, Alessandro Pedrani, Andrea Caciolai, Kay Rottmann, and Davide Bernardi. 2023. A survey on knowledge editing of neural networks. *CoRR*, abs/2310.19704.

Nat McAleese, Rai (Michael Pokorny), and Juan Felipe Cerón Uribe. 2024. Llm critics help catch llm bugs.

Thomas McGrath, Andrei Kapishnikov, Nenad Tomasev, Adam Pearce, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. 2021. Acquisition of chess knowledge in alphazero. *CoRR*, abs/2111.09259.

Karen L. McGraw and Karan Harbison-Briggs. 1990. *Knowledge acquisition - principles and guidelines*. Prentice Hall.

Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. 2024. Shortgpt: Layers in large language models are more redundant than you expect. *CoRR*, abs/2403.03853.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *CoRR*, abs/2405.14734.

William Merrill, Nikolaos Tsilivis, and Aman Shukla. 2023. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. *CoRR*, abs/2303.11873.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2023a. Circuit component reuse across tasks in transformer language models. *CoRR*, abs/2310.08744.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2023b. A mechanism for solving relational tasks in transformer language models.

Joseph Miller and Clement Neo. 2024. We found an neuron in gpt-2.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ali Momeni, Babak Rahmani, Benjamin Scellier, Logan G. Wright, Peter L. McMahon, Clara C. Wanjura, Yuhang Li, Anas Skalli, Natalia G. Berloff, Tatsuhiro Onodera, Ilker Oguz, Francesco Morichetti, Philipp del Hougne, Manuel Le Gallo, Abu Sebastian, Azalia Mirhoseini, Cheng Zhang, Danijela Marković, Daniel Brunner, Christophe Moser, Sylvain Gigan, Florian Marquardt, Aydogan Ozcan, Julie Grollier, Andrea J. Liu, Demetri Psaltis, Andrea Alù, and Romain Fleury. 2024. Training of physical neural networks. *arXiv preprint arXiv:2406.03372*.

Ari S. Morcos, Haonan Yu, Michela Paganini, and Yuandong Tian. 2019. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4933–4943.

Dan Mossing, Steven Bills, Henk Tillman, Tom Dupré la Tour, Nick Cammarata, Leo Gao, Joshua Achiam, Catherine Yeh, Jan Leike, Jeff Wu, and William Saunders. 2024. Transformer debugger. `https://github.com/openai/transformer-debugger`.

Théo Moutakanni, Piotr Bojanowski, Guillaume Chassagnon, Céline Hudelot, Armand Joulin, Yann LeCun, Matthew Muckley, Maxime Oquab, Marie-Pierre Revel, and Maria Vakalopoulou. 2024. Advancing human-centric AI for robust x-ray analysis through holistic self-supervised learning. *CoRR*, abs/2405.01469.

Anirban Mukherjee and Hannah Hanwen Chang. 2024. AI knowledge and reasoning: Emulating expert creativity in scientific research. *CoRR*, abs/2404.04436.

Alhassan Mumuni, Fuseini Mumuni, and Nana Kobina Gerrar. 2024. A survey of synthetic data augmentation methods in computer vision. *CoRR*, abs/2403.10075.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Interpretable machine learning: definitions, methods, and applications. *CoRR*, abs/1901.04592.

Jatin Nainani. 2024. Evaluating brain-inspired modular training in automated circuit discovery for mechanistic interpretability. *CoRR*, abs/2401.03646.

Neel Nanda. 2023. How to think about activation patching.

Neel Nanda and Joseph Bloom. 2022. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *CoRR*, abs/2307.06435.

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. 2024. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models. *CoRR*, abs/2406.02061.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*.

Shiwen Ni, Dingwei Chen, Chengming Li, Xiping Hu, Ruifeng Xu, and Min Yang. 2023. Forgetting before learning: Utilizing parametric arithmetic for knowledge updating in large language models. *CoRR*, abs/2311.08011.

Aline Normoyle, João Sedoc, and Funda Durupinar. 2024. Using llms to animate interactive story characters with emotions and personality. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, VR Workshops 2024, Orlando, FL, USA, March 16-21, 2024*, pages 632–635. IEEE.

nostalgebraist. 2020. interpreting gpt: the logit lens.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*.

Bruno A Olshausen and David J Field. 1997. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *CoRR*, abs/2209.11895.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in llms.

Marina A Pak, Karina A Markhieva, Mariia S Novikova, Dmitry S Petrov, Ilya S Vorobyev, Ekaterina S Maksimova, Fyodor A Kondrashov, and Dmitry N Ivankov. 2023. Using alphafold to predict the impact of single mutations on protein stability and function. *Plos one*, 18(3):e0282689.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C. Wallace, and David Bau. 2023. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning, CoNLL 2023, Singapore, December 6-7, 2023*, pages 548–560. Association for Computational Linguistics.

Haowen Pan, Yixin Cao, Xiaozhi Wang, and Xun Yang. 2023a. Finding and editing multi-modal neurons in pre-trained transformer. *CoRR*, abs/2311.07470.

Haowen Pan, Yixin Cao, Xiaozhi Wang, Xun Yang, and Meng Wang. 2024a. Finding and editing multi-modal neurons in pre-trained transformers.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023b. Large language models and knowledge graphs: Opportunities and challenges.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024b. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023a. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023b. The linear representation hypothesis and the geometry of large language models. *CoRR*, abs/2311.03658.

Adam Pearce, Asma Ghandeharioun, Nada Hussein, Nithum Thain, Martin Wattenberg, and Lucas Dixon. 2023. Do machine learning models memorize or generalize?

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan S. Wind, Stanislaw Wozniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. RWKV: reinventing rnns for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 14048–14077. Association for Computational Linguistics.

Roger Penrose. Limitations of the discrete.

Antoine Bellemare Pépin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. 2024. Divergent creativity in humans and large language models. *CoRR*, abs/2405.13012.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

P Jonathon Phillips, P Jonathon Phillips, Carina A Hahn, Peter C Fontana, Amy N Yates, Kristen Greene, David A Broniatowski, and Mark A Przybocki. 2021. Four principles of explainable artificial intelligence.

USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir S V, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A. Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, and Naomi Saphra. 2024. Recite, reconstruct, recollect: Memorization in lms as a multifaceted phenomenon. *CoRR*.

Penghui Qi, Xinyi Wan, Guangxing Huang, and Min Lin. 2024. Zero bubble pipeline parallelism. *CoRR*, abs/2401.10241.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *CoRR*, abs/2307.07924.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2024a. Scaling large-language-model-based multi-agent collaboration. *arxiv.org/abs/2406.07155*.

Chen Qian, Jie Zhang, Wei Yao, Dongrui Liu, Zhenfei Yin, Yu Qiao, Yong Liu, and Jing Shao. 2024b. Towards tracing trustworthiness dynamics: Revisiting pre-training period of large language models. *CoRR*, abs/2402.19465.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5368–5393. Association for Computational Linguistics.

Shuofei Qiao, Ningyu Zhang, Runnan Fang, Yujie Luo, Wangchunshu Zhou, Yuchen Eleanor Jiang, Chengfei Lv, and Huajun Chen. 2024. AUTOACT: automatic agent learning from scratch via self-planning. *CoRR*, abs/2401.05268.

Jiaxin Qin, Zixuan Zhang, Chi Han, Manling Li, Pengfei Yu, and Heng Ji. 2024. Why does new knowledge create messy ripple effects in llms? *CoRR*, abs/2407.12828.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024a. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024b. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

Ralph Raiola. 2023. Chatgpt, can you tell me a story? an exercise in challenging the true creativity of generative AI. *Commun. ACM*, 66(5).

Harsh Raj, Domenic Rosati, and Subhabrata Majumdar. 2022. Measuring reliability of large language models through semantic consistency. *CoRR*, abs/2211.05853.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2023, Raleigh, NC, USA, February 8-10, 2023*, pages 464–483. IEEE.

Jing Ren and Feng Xia. 2024. Brain-inspired artificial intelligence: A comprehensive review. *arXiv preprint arXiv:2408.14811*.

Mengjie Ren, Boxi Cao, Hongyu Lin, Cao Liu, Xianpei Han, Ke Zeng, Guanglu Wan, Xunliang Cai, and Le Sun. 2024a. Learning or self-aligning? rethinking instruction fine-tuning.

Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan H Kim, et al. 2024b. Safetywashing: Do ai safety benchmarks actually measure safety progress? *arXiv preprint arXiv:2407.21792*.

Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, et al. 2024. Open problems in technical ai governance. *arXiv preprint arXiv:2407.14981*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5418–5426. Association for Computational Linguistics.

Mark A Runco and Garrett J Jaeger. 2012. The standard definition of creativity. *Creativity research journal*, 24(1):92–96.

Solve Sæbø and Helge Brovold. 2024. On the stochastics of human and artificial creativity. *CoRR*, abs/2403.06996.

Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, George Gor, Emma Bluemke, Sarah Shoker, Janet Egan, Robert F. Trager, Shahar Avin, Adrian Weller, Yoshua Bengio, and Diane Coyle. 2024. Computing power and the governance of artificial intelligence. *CoRR*, abs/2402.08797.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. 2024. Rethinking LLM memorization through the lens of adversarial compression. *CoRR*, abs/2404.15146.

Sarah Schwettmann, Neil Chowdhury, Samuel Klein, David Bau, and Antonio Torralba. 2023a. Multimodal neurons in pretrained text-only transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 2854–2859. IEEE.

Sarah Schwettmann, Tamar Rott Shaham, Joanna Materzynska, Neil Chowdhury, Shuang Li, Jacob Andreas, David Bau, and Antonio Torralba. 2023b. FIND: A function description benchmark for evaluating interpretability methods. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Rémi Segretain, Sergiu Ivanov, Laurent Trilling, and Nicolas Glade. 2020. A methodology for evaluating the extensibility of boolean networks' structure and function. In *Complex Networks & Their Applications IX - Volume 2, Proceedings of the Ninth International Conference on Complex Networks and Their Applications, COMPLEX NETWORKS 2020, 1-3 December 2020, Madrid, Spain*, volume 944 of *Studies in Computational Intelligence*, pages 372–385. Springer.

Raj Sanjay Shah, Khushi Bhardwaj, and Sashank Varma. 2024. Development of cognitive intelligence in pre-trained language models. *arXiv preprint arXiv:2407.01047*.

Jingbo Shang, Zai Zheng, Xiang Ying, Felix Tao, and Mindverse Team. 2024. Ai-native memory: A pathway from llms towards agi. *arXiv preprint arXiv:2406.18312*.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 2257–2273. Association for Computational Linguistics.

Lee Sharkey, Dan Braun, , and Beren Millidge. 2022. Taking features out of superposition with sparse autoencoders.

Arnab Sen Sharma, David Atkinson, and David Bau. 2024. Locating and editing factual associations in mamba. *CoRR*, abs/2404.03646.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *CoRR*, abs/2305.14739.

Jung-Eun Shin, Adam J. Riesselman, Aaron W. Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C. Kruse, and Debora S. Marks. 2021. Protein design and variant prediction using autoregressive generative models. *Nature Communications*, 12(1).

Ravid Shwartz-Ziv and Yann LeCun. 2024. To compress or not to compress - self-supervised learning and information theory: A review. *Entropy*, 26(3):252.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *CoRR*, abs/2402.01761.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek

Natarajan. 2022. Large language models encode clinical knowledge. *CoRR*, abs/2212.13138.

Andries P. Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going mad? A look at multi-agent debate strategies for llms. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org.

Andrea Soltoggio, Eseoghene Ben-Iwhiwhu, Vladimir Braverman, Eric Eaton, Benjamin Epstein, Yunhao Ge, Lucy Halperin, Jonathan How, Laurent Itti, Michael A Jacobs, et al. 2024. A collective ai via lifelong learning and sharing at the edge. *Nature Machine Intelligence*, 6(3):251–264.

Spivey and Michael. 2007. The continuity of mind.

Larry Squire, Darwin Berg, Floyd E Bloom, Sascha Du Lac, Anirvan Ghosh, and Nicholas C Spitzer. 2012. *Fundamental neuroscience*. Academic press.

Michael Stenger, Robert Leppich, Ian T. Foster, Samuel Kounev, and André Bauer. 2024. Evaluation is key: a survey on evaluation measures for synthetic time series. *J. Big Data*, 11(1):66.

Robert J Sternberg. 2006. The nature of creativity. *Creativity research journal*, 18(1):87.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7035–7052. Association for Computational Linguistics.

Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. 2024. Confidence regulation neurons in language models. *arXiv preprint arXiv:2406.16254*.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. 2023. Getting aligned on representational alignment. *CoRR*, abs/2310.13018.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. 2024. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv:2407.04620*.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive network: A successor to transformer for large language models. *CoRR*, abs/2307.08621.

Mujeen Sung, Jinhyuk Lee, Sean S. Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4723–4734. Association for Computational Linguistics.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent LLM agents. *CoRR*, abs/2306.03314.

Alex Tamkin, Mohammad Taufeeque, and Noah D. Goodman. 2023. Codebook features: Sparse and discrete interpretability for neural networks. *CoRR*, abs/2310.17230.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *CoRR*, abs/2402.16438.

Bozhong Tian, Xiaozhuan Liang, Siyuan Cheng, Qingbin Liu, Mengru Wang, Dianbo Sui, Xi Chen, Huajun Chen, and Ningyu Zhang. 2024. To forget or not? towards practical knowledge unlearning for large language models. *arXiv preprint arXiv:2407.01920*.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2023. Function vectors in large language models. *CoRR*, abs/2310.15213.

Tom Tseng, Euan McLean, Kellin Pelrine, Tony T. Wang, and Adam Gleave. 2024. Can go ais be adversarially robust? *arXiv preprint arXiv:2406.12843*.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *CoRR*, abs/2308.10248.

Vashishtha, Aniket, Kumar, Abhinav, Reddy, Abbavaram Gowtham, Balasubramanian, Vineeth N, Sharma, and Amit. 2024. Teaching transformers causal reasoning through axiomatic training. *arXiv preprint arXiv:2407.07612*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Martina G. Vilas, Federico Adolfi, David Poeppel, and Gemma Roig. 2024. Position paper: An inner interpretability framework for ai inspired by lessons from cognitive neuroscience. *CoRR*, abs/2406.01352.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. 2024a. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. *CoRR*, abs/2405.15071.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *CoRR*, abs/2310.07521.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023b. A causal view of entity bias in (large) language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15173–15184. Association for Computational Linguistics.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023c. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024b. Detoxifying large language models via knowledge editing. *CoRR*, abs/2403.14472.

Peng Wang, Xiang Wei, Fangxu Hu, and Wenjuan Han. 2024c. Transgpt: Multi-modal generative pre-trained transformer for transportation. *CoRR*, abs/2402.07233.

Ruoyao Wang, Graham Todd, Ziang Xiao, Xingdi Yuan, Marc-Alexandre Côté, Peter Clark, and Peter Jansen. 2024d. Can language models serve as text-based world simulators? *CoRR*, abs/2406.06485.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2023d. Knowledge editing for large language models: A survey. *CoRR*, abs/2310.16218.

Xiaohan Wang, Shengyu Mao, Ningyu Zhang, Shumin Deng, Yunzhi Yao, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024e. Editing conceptual knowledge for large language models. *CoRR*, abs/2403.06259.

Yuanfei Wang, Fangwei Zhong, Jing Xu, and Yizhou Wang. 2022. Tom2c: Target-oriented multi-agent communication and cooperation with theory of mind. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2023. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *CoRR*, abs/2308.09729.

Lilian Weng. 2024. Extrinsic hallucinations in llms.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4602–4625. Association for Computational Linguistics.

Leslie Owen Wilson. 2016. Anderson and krathwohl–bloom's taxonomy revised.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023a. DEPN: detecting and editing privacy neurons in pretrained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2875–2886. Association for Computational Linguistics.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023b. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *CoRR*, abs/2307.02477.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2024. Reft: Representation finetuning for language models. *CoRR*, abs/2404.03592.

Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Dingwen Yang, Chenyang Liao, Xin Guo, Wei He, Songyang Gao, Lu Chen, Rui Zheng, Yicheng Zou, Tao Gui,

Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. 2024. Agentgym: Evolving large language model-based agents across diverse environments.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip H. S. Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behaviors? *CoRR*, abs/2402.04559.

Fangzhi Xu, Qiushi Sun, Kanzhi Cheng, Jun Liu, Yu Qiao, and Zhiyong Wu. 2024a. Interactive evolution: A neural-symbolic self-training framework for large language models.

Rongwu Xu, Zehan Qi, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. Knowledge conflicts for llms: A survey. *CoRR*, abs/2403.08319.

Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. 2024c. AI for social science and social science of AI: A survey. *CoRR*, abs/2401.11839.

Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. 2024d. Hallucination is inevitable: An innate limitation of large language models. *CoRR*, abs/2401.11817.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jianhao Yan, Futing Wang, Yafu Li, and Yue Zhang. 2024. Potential and challenges of model editing for social debiasing. *CoRR*, abs/2402.13462.

Diji Yang, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Jie Yang, and Yi Zhang. 2024a. IM-RAG: multi-round retrieval-augmented generation through learning inner monologues. *CoRR*, abs/2405.13021.

Hongkang Yang, Zehao Lin, Wenjin Wang, Hao Wu, Zhiyu Li, Bo Tang, Wenqiang Wei, Jinbo Wang, Zeyun Tang, Shichao Song, Chenyang Xi, Yu Yu, Kai Chen, Feiyu Xiong, Linpeng Tang, and Weinan E. 2024b. Memory^3: Language modeling with explicit memory. *arxiv.org/abs/2407.01178*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models.

Wanli Yang, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, and Xueqi Cheng. 2024c. The butterfly effect of model editing: Few edits can trigger large language models collapse. *CoRR*, abs/2402.09656.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023a. Large language model unlearning. *CoRR*, abs/2310.10683.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023b. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240, Singapore. Association for Computational Linguistics.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. *CoRR*, abs/2405.17969.

Tian Ye, Zicheng Xum, Yuanzhi Li, and Zeyuan Allen-Zhu. 2024. Physics of language models: Part 2.1, grade-school math and the hidden reasoning process. *arXiv preprint arXiv:2407.20311*.

Huizi Yu, Lizhou Fan, Lingyao Li, Jiayan Zhou, Zihui Ma, Lu Xian, Wenyue Hua, Sijia He, Mingyu Jin, Yongfeng Zhang, Ashvin Gandhi, and Xin Ma. 2024a. Large language models in biomedical and health informatics: A bibliometric review.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Nianyi Lin, Kaifeng Yun, Linlu Gong, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, and Juanzi Li. 2023a. Kola: Carefully benchmarking world knowledge of large language models. *CoRR*, abs/2306.09296.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2023b. Language models are super mario: Absorbing abilities from homologous models as a free lunch. *CoRR*, abs/2311.03099.

Lei Yu, Meng Cao, Jackie Chi Kit Cheung, and Yue Dong. 2024b. Mechanisms of non-factual hallucinations in language models. *CoRR*, abs/2403.18167.

Lei Yu, Jingcheng Niu, Zining Zhu, and Gerald Penn. 2024c. Functional faithfulness in the wild: Circuit discovery with differentiable computation graph pruning. *arXiv preprint arXiv:2407.03779*.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023c. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9924–9959. Association for Computational Linguistics.

Zeyu Yun, Yubei Chen, Bruno A. Olshausen, and Yann LeCun. 2021. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In *Proceedings of Deep Learning Inside Out: The 2nd Workshop on Knowledge Extraction and Integration for Deep*

Learning Architectures, DeeLIO@NAACL-HLT 2021, Online, June 10 2021, pages 1–10. Association for Computational Linguistics.

Linda Zagzebski. 2017. What is knowledge? *The Blackwell guide to epistemology*, pages 92–116.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Fred Zhang and Neel Nanda. 2023. Towards best practices of activation patching in language models: Metrics and methods. *CoRR*, abs/2309.16042.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2023a. Exploring collaboration mechanisms for LLM agents: A social psychology view. *CoRR*, abs/2310.02124.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023b. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 794–812. Association for Computational Linguistics.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. 2024a. A comprehensive study of knowledge editing for large language models. *CoRR*, abs/2401.01286.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023c. Instruction tuning for large language models: A survey. *CoRR*, abs/2308.10792.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Yipeng Zhang, Haitao Mi, and Helen Meng. 2024b. Self-tuning: Instructing llms to effectively acquire new knowledge through self-teaching.

Yuji Zhang, Sha Li, Jiateng Liu, Pengfei Yu, Yi R Fung, Jing Li, Manling Li, and Heng Ji. 2024c. Knowledge overshadowing causes amalgamated hallucination in large language models. *arXiv preprint arXiv:2407.08039*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics.

Zhexin Zhang, Junxiao Yang, Pei Ke, and Minlie Huang. 2023d. Defending large language models against jailbreaking attacks through goal prioritization. *CoRR*, abs/2311.09096.

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, and Zhi-Qin John Xu. 2024d. Initialization is critical to whether transformers fit composite functions by inference or memorizing. *CoRR*, abs/2405.05409.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024a. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2):20:1–20:38.

Jun Zhao, Zhihao Zhang, Yide Ma, Qi Zhang, Tao Gui, Luhui Gao, and Xuanjing Huang. 2023a. Unveiling A core linguistic region in large language models. *CoRR*, abs/2310.14928.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024b. Retrieval-augmented generation for ai-generated content: A survey. *CoRR*, abs/2402.19473.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023b. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Chujie Zheng, Ziqi Wang, Heng Ji, Minlie Huang, and Nanyun Peng. 2024. Weak-to-strong extrapolation expedites alignment. *CoRR*, abs/2404.16792.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5017–5033. Association for Computational Linguistics.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15686–15702. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. LIMA: less is more for alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yilun Zhou, Caiming Xiong, Silvio Savarese, and Chien-Sheng Wu. 2024. Shared imagination: Llms hallucinate alike. *arXiv preprint arXiv:2407.16604*.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. 2023b. How alignment and jailbreak work: Explain llm safety through intermediate hidden states. *CoRR*, abs/2309.04827.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. 2024. Language models represent beliefs of self and others. *CoRR*, abs/2402.18496.

Zeyuan Allen Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *CoRR*, abs/2309.14316.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. 2024. Improving alignment and robustness with circuit breakers. *CoRR*, abs/2406.04313.

A Preliminary

A.1 Knowledge Scope

We denote a diverse array of knowledge as set $\mathbf{K}$, wherein each element $k \in \mathbf{K}$ is a specific piece of knowledge, which can be expressed by various records, e.g., a text record "*The president of the United States in 2024 is Biden*" (denoted as r_k).

A.2 Definition of Knowledge in LLMs

Given a LLM denoted as $\mathcal{F}$, we formulate that $\mathcal{F}$ master knowledge k if $\mathcal{F}$ can correctly answer the corresponding question $r_{k\backslash t}$:

$$\begin{aligned} &t = \mathcal{F}\left(r_{k\backslash t}\right) \\ &(t \in \mathbf{T}) \Rightarrow (\mathcal{F} \text{ masters knowledge } k), \end{aligned} \tag{1}$$

t is the output of a LLM $\mathcal{F}$, $r_{k\backslash t}$ is a record about knowledge k that lacks pivot information. Take an example for illustration: $r_{k\backslash t}$ is *"The president of the United States in 2024 is __"*, the pivot information is "Biden". Note that, $r_{k\backslash t}$ can be represented by the above textual statement, captured through a question-answering pair ("*Who is the President of the United States in 2024?*", or conveyed by audio, video, image [5], and other equivalent expressions. The pivot information for $r_{k\backslash t}$ can be expressed by various fomarts, which are formulated as $\mathbf{T} = \{\text{"Biden"}, \text{"Joe Biden"}, \cdots\}$. If the output t is an element from the correct answer set $\mathbf{T}$, we hypothesize that $\mathcal{F}$ master knowledge k.

A.3 The Architecture of LLMs

An LLM $\mathcal{F}$ consists of numerous neurons, which work systematically under a specific architecture.

Transformer-based architecture. The prevailing architecture in current LLMs is the Transformer (Vaswani et al., 2017). Specifically, a transformer-based LLM $\mathcal{F}$ begins with a token embedding, followed by L layers transformer block, and ends with token unembedding used for predicting answer tokens. Each transformer block layer l consists of Attention Heads (Attention) and Multilayer Perceptron (MLP):

$$h_{l+1} = h_l + \text{MLP}\left(h_l + \text{Attention}\left(h_l\right)\right), \tag{2}$$

h_l is the hidden state from l-th layer.

Other architectures. Other architectures including competitive variants of the transformer, e.g., SSM (Gu and Dao, 2023), TTT (Sun et al., 2024) and RWKV (Peng et al., 2023), and architectures in computer vision (Li et al., 2023c) and multi-modal fields are detailed in §C.1.

A.4 Knowledge Analysis Methods

Knowledge analysis method $\mathcal{M}$ aims to interpret how LLMs work inside and reveal precise causal connections between specific components and outputs (Bereska and Gavves, 2024). Furthermore, if components $\mathbf{C}$ of $\mathcal{F}$ accurately infer t through analysis method $\mathcal{M}$, it is assumed that the knowledge k is presented by $\mathbf{C}$:

$$\begin{aligned} &t = \mathcal{M}_{\mathbf{C}\subseteq\mathcal{F}}\left(r_{k\backslash t}, \mathbf{C}\right), \\ &(t \in \mathbf{T}) \Rightarrow (\mathbf{C} \text{ represents knowledge } k), \end{aligned} \tag{3}$$

The elements in set $\mathbf{C}$ may be individual neurons, MLPs, attention heads, a transformer block layer, or knowledge circuit (Yao et al., 2024). These methods are divided into two categories: observation and intervention (Bereska and Gavves, 2024).

[5] While audio, video, and image records have been somewhat investigated, they are still relatively unexplored areas and thus are only discussed in §G.2 and §C.2.

Observation-based methods. These methods aim to observe the internal information of $\mathcal{F}$, directly projecting the output of component $\mathbf{C}$ into human-understandable forms by E:

$$t = E_{\mathbf{C}\subseteq\mathcal{F}}\left(r_{k\setminus t}, \mathbf{C}, \mathcal{F}\right), \tag{4}$$

E is a evaluation metric, which can be a probe (Räuker et al., 2023), logit lens (nostalgebraist, 2020), or a sparse representation (Gao et al., 2024c). **Probe** is a meticulously trained classifier, and its classification performance is used to observe the relationship between model's behavior and the output of $\mathbf{C}$ (Belinkov, 2022; Elazar et al., 2021; McGrath et al., 2021; Gurnee et al., 2023). **Logit lens** usually translate output of $\mathbf{C}$ into vocabulary tokens via token unembedding (Geva et al., 2022b; Belrose et al., 2023; Pal et al., 2023; Din et al., 2024; Langedijk et al., 2023). **Sparse representation** maps the output of $\mathbf{C}$ into a higher-dimensional space with strong sparsity through dictionary learning (He et al., 2024b; Olshausen and Field, 1997; Yun et al., 2021; Karvonen et al., 2024), with sparse auto-encoder (Sharkey et al., 2022; Cunningham et al., 2023; Lee et al., 2006; Gao et al., 2024a) being a prominent example. The higher-dimensional space represents independent (or monosemantic (Bricken et al., 2023)) and interpretable features more easily (Rai et al., 2024b). The output of $\mathbf{C}$ is the combination (Elhage et al., 2022; Bricken et al., 2023) of these features.

Intervention-based methods. These methods allow for direct corruptions in LLMs to identify the critical $\mathbf{C}$ via intervention strategies $\mathcal{I}$. Note that $\mathbf{C}$, encompassing various neuron combinations, correlates with specific model behaviors:

$$\begin{aligned} \mathbf{C} &= \mathcal{I}\left(r_{k\setminus t}, \mathcal{F}\right), \\ t &= E\left(r_{k\setminus t}, \mathbf{C}, \mathcal{F}\right) \end{aligned} \tag{5}$$

$\mathcal{I}$ is also known as causal mediation analysis (Vig et al., 2020), causal tracing (Meng et al., 2022), interchange interventions (Geiger et al., 2022), activation patching (Wang et al., 2023c; Zhang and Nanda, 2023), path patching (Goldowsky-Dill et al., 2023), and causal scrubbing techniques (LawrenceC et al., 2022). Specifically, $\mathcal{I}$ consists of the following three steps. 1) **Clean run**: $\mathcal{F}$ generates the correct answer t based on the input $r_{k\setminus t}$. 2) **Corrupted run**: corrupt the generation process of $\mathcal{F}$ in the *clean run* by introducing noise into the input or neurons (Meng et al., 2022; Goldowsky-Dill et al., 2023; Stolfo et al., 2023; Yao et al., 2024; Conmy et al., 2023; Mossing et al., 2024; Lepori et al., 2023; Huang et al., 2023a). 3) **Restoration run**: recover the correct answer t by restoring unnoised information from $\mathbf{C}$ (Meng et al., 2022; Vig et al., 2020; Wang et al., 2023c; Zhang et al., 2017; Nanda, 2023). For intervention-based methods, E typically refers to the token unembedding used for predicting answer tokens. Under the evaluation metric E, there exists a causal relationship between $\mathbf{C}$ and specific behavior of LLMs $\mathcal{F}$ in Eq 5.

Based on the aforementioned preliminary discussion, the taxonomy of knowledge mechanisms in LLMs is illustrated in Fig 4.

B Comparison of Different Analysis Methods

B.1 Comparison of Different Mechanism Analysis Methods

The above four Hypotheses are achieved by Observation-based and Intervention-based methods. These two methods are typically combined to trace knowledge in LLMs (Mossing et al., 2024; Ghandeharioun et al., 2024). Most knowledge analysis methods are architecture-agnostic and can be adapted to various models.

Each method is suitable for different scenarios. Specifically, the Modular Region Hypothesis can be analyzed using either Observation-based or Intervention-based methods. In contrast, the Connection Hypothesis, which examines inter-regional connectivity, generally necessitates Intervention-based methods. However, the results of knowledge mechanism analysis depend heavily on different methods and are sensitive to evaluation metrics and implementation details (Schwettmann et al., 2023b). Hence, Huang et al. (2024b) propose a dataset, RAVEL, to quantify the comparisons between a variety of existing interpretability methods. They suggest that methods with supervision are better than methods with unsupervised featurizers. Later, Zhang and Nanda (2023) further systematically examine the impact of methodological details in intervention-based methods. For corrupted run, they recommend Symmetric Token Replacement (e.g., "The Eiffel Tower"→"The Colosseum") (Sharma et al., 2024; Vig et al., 2020) instead of Gaussian Noising (Meng et al., 2022), which disrupts the model's internal mechanisms. For metric E, both logit lens and probe can be employed to trace factual knowledge (Meng et al., 2022), where the target output is typically few tokens. In this

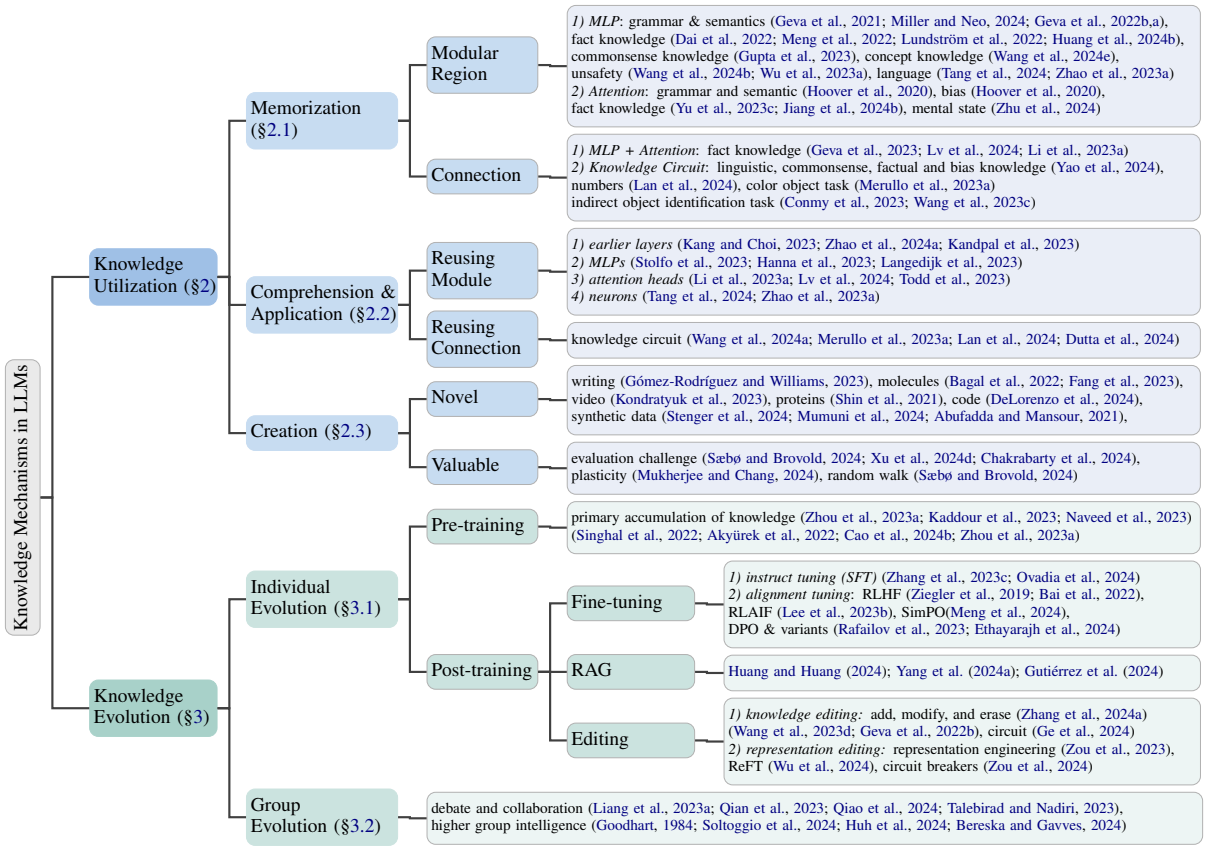

Figure 4: The taxonomy of knowledge mechanisms in LLMs.

scenario, Zhang and Nanda (2023) advocate using the logit lens over probes for evaluation metric E due to its fine-grained control over localization outcomes. Moreover, probe is capable of exploring abstract knowledge and abilities, such as theory of mind or mental states (Zhu et al., 2024; Ye et al., 2024; Jin, 2024), where the target output requires multiple tokens to express. Jin (2024) suggest that deeper probes are more (generally) more accurate.

B.2 Comparison of Different Evolution Strategies

Individuals and groups achieve dynamic intelligence primarily through two strategies: updating internal parametric knowledge (Zhou et al., 2023a; Qiao et al., 2024) and leveraging external knowledge [6] (Huang and Huang, 2024; Xie et al., 2024). These two strategies are usually used together in applications (Yang et al., 2024b).

Updating internal parametric knowledge necessitates high-quality data for parameter adjustments (Vashishtha et al., 2024; Cao et al., 2024a). Data proves pivotal when fine-tuning models to acquire new knowledge. Ovadia et al. (2024) also posit that the continued training of LLMs via unsupervised tuning generally exhibits suboptimal performance when it comes to acquiring new knowledge. Note that updating internal parametric knowledge requires resolving conflicts among internal parameters. The crux of effective internal knowledge updating lies in preserving the consistency of the model's parameter knowledge before and after tuning. In contrast, *leveraging external knowledge* requires managing conflicts within the external knowledge itself [7] as well as conflicts between external and internal knowledge (Xu et al., 2024b; Liu et al., 2024a). Besides, parametric knowledge compresses extensive information, promoting grokking and enhancing generalization (Wang et al., 2024a). In contrast, leveraging external knowledge avoids high training costs but necessitates substantial maintenance and retrieval costs for every user query. Therefore, the *combination of these two strategies* is promising. An attempt for combination (Yang et al., 2024b) suggests em-

[6] Leveraging external knowledge includes using prompts (Xie et al., 2024), ICL, and RAG.

[7] Inconsistencies in external information are common, as external documents often contain conflicting data, particularly in contexts for RAG.

ploying RAG for low-frequency knowledge and parametric strategy for high-frequency knowledge.

B.2.1 Comparison of Methods for Knowledge Evolution

Note that due to the page limit, §3 does not provide a detailed enumeration of various techniques and details, such as machine unlearning and knowledge augmentation. Hence, we briefly outline common methods during post-training stage in this section and illustrate their associations and differences (Zhang et al., 2024a) in Fig 5.

- *Continual Learning* aims to continually acquire new skills and learn new tasks while retaining previously acquired knowledge.
- *Parameter-efficient Fine-tuning* (PET) (Zhang et al., 2019) only updates a minimal set of parameters instead of full fine-tuning. A promising strategy is LoRA (Hu et al., 2022).
- *Knowledge Augmentation* is proposed to assist the model in handling unknown knowledge for LLMs (Zhang et al., 2019; Han et al., 2022). RAG (Huang and Huang, 2024) is the is the most prevalent methods. Beside, knowledge augmentation also includes prompt engineering (Gu et al., 2023; Kraljic and Lahav, 2024; Liang et al., 2023b) and in-context learning (Luo et al., 2024a).
- *Machine Unlearning* (Nguyen et al., 2022; Tian et al., 2024; Liu et al., 2024d) focuses on discarding undesirable behaviors from LLMs.
- *Editing*, including knowledge editing (Zhang et al., 2024a) and representation editing (Wu et al., 2024), aims to enable quick and precise modifications to the LLMs. Usually, editing first identifies the knowledge location in LLMs and then precisely modifies model behavior through a few instances.

C Universality Intelligence

To validate the hypotheses in this paper across different architectures, we first introduce other popular model architectures in §C.1, and observe the generalizability of our hypotheses across other model architectures in §C.2. Besides, recent work further claims that models trained with different data, modalities, and objectives are converging to shared representation spaces (Huh et al., 2024). Artificial and biological neural networks [8] also share

[8]Unless otherwise specified as biological networks, the terms models, neural networks, neural models, and machines refer to artificial neural networks.

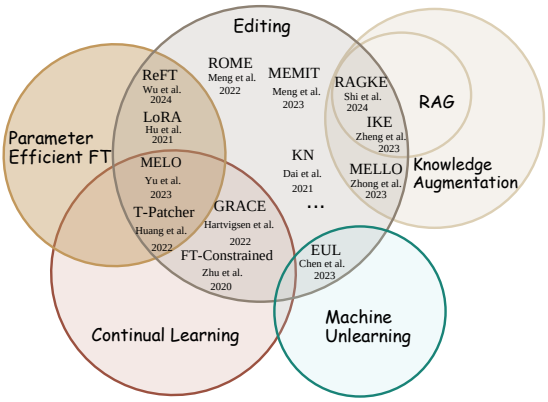

Figure 5: Comparison of Different Methods for Knowledge Evolution.

similar features and circuits, suggesting a universal underlying mechanism (Sucholutsky et al., 2023; Chan et al., 2023; Kornblith et al., 2019). Therefore, analogous to biological taxonomy, we introduce artificial neural model family and discuss the potential universality intelligence in the future in §C.3.

C.1 Model Architecture

C.1.1 Transformer

MLP The Multilayer Perceptron (MLP) is a crucial component in neural networks, usually comprising multiple fully connected layers. Within the Transformer architecture, the MLP plays a vital role in applying nonlinear transformations to the input hidden states, thereby enriching the model's capacity for expression. More precisely, every MLP block involves two linear transformations separated by a point-wise activation function σ:

$$\mathrm{MLP}^l\left(h^l\right) = \sigma\left(W_K^l h^l\right) W_V^l, \tag{6}$$

where σ is the point-wise activation function, typically a non-linear function such as ReLU or GELU. W_K^l is the weight matrix for the first linear transformation in the l-th layer, mapping the input hidden state h^l to an intermediate representation. W_V^l is the weight matrix for the second linear transformation in the l-th layer, transforming the intermediate representation to the output of the MLP block.

Attention is a mechanism in neural networks, especially in models like Transformers, that captures dependencies between different positions within a sequence. It works by transforming each input element into Query (Q), Key (K), and Value (V) vectors, computing attention scores between elements, and then calculating a weighted sum of val-

ues based on these scores. Specifically, for an input sequence represented as matrix X, the transformations are as follows:

$$\begin{aligned} Q &= XW^Q, \\ K &= XW^K, \\ V &= XW^V, \end{aligned} \tag{7}$$

where W^Q, W^K, and W^V are learned projection matrices. The attention scores are computed using the scaled dot-product attention mechanism:

$$H = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{8}$$

where d_k is the dimensionality of the Key vectors. This allows the model to focus on different parts of the sequence adaptively, making it effective for tasks like natural language processing where understanding long-range dependencies is important.

Variants of Transformer Variants of the Transformer also achieve success. For instance, RWKV (Peng et al., 2023) combines the efficient parallelizable training of transformers with the efficient inference of RNNs while mitigating their limitations. TTT (Sun et al., 2024) replaces the hidden state of an RNN with a machine learning model. TTT compresses context through actual gradient descent on input tokens. RetNet (Sun et al., 2023) theoretically derives the connection between recurrence and attention, simultaneously achieves training parallelism, low-cost inference, and good performance.

C.1.2 SSM

Mamba introduced by Gu and Dao (2023), is a recent family of autoregressive language models based on state space models (SSMs). Mamba employs a unique architecture called MambaBlock, which replaces the attention and MLP blocks used in Transformer layers.

Specifically, Mamba maps a sequence of tokens $x = [x_1, x_2, \ldots, x_T]$ to a probability distribution over the next token y. Each token x_i is first embedded into a hidden state of size d as $h_i^{(0)}$, which is then transformed sequentially by a series of MambaBlocks. The hidden state $h_i^{(\ell)}$ after the ℓ-th MambaBlock is computed as follows:

$$h_i^{(\ell)} = h_i^{(\ell-1)} + o_i^{(\ell)} \tag{9}$$

The output $o_i^{(\ell)}$ of the ℓ-th MambaBlock for the i-th token is a combination of $s_i^{(\ell)}$ (from Conv and SSM operations) and $g_i^{(\ell)}$ (a gating mechanism):

$$\begin{aligned} o_i^{(\ell)} &= \text{MambaBlock}^{(\ell)}\left[h_1^{(\ell-1)}, h_2^{(\ell-1)}, \ldots, h_i^{(\ell-1)}\right] \\ &= W_o^{(\ell)}\left[s_i^{(\ell)} \otimes g_i^{(\ell)}\right] \end{aligned} \tag{10}$$

Here, $\otimes$ denotes element-wise multiplication. The calculation of $s_i^{(\ell)}$ is as follows:

$$a_i^{(\ell)} = W_a^{(\ell)} h_i^{(\ell)} \tag{11}$$

$$\begin{aligned} &c_1^{(\ell)}, c_2^{(\ell)}, \ldots, c_i^{(\ell)} = \\ &\text{SiLU}\left[\text{Conv1D}\left[a_1^{(\ell)}, a_2^{(\ell)}, \ldots, a_i^{(\ell)}\right]\right] \end{aligned} \tag{12}$$

$$s_i^{(\ell)} = \text{selective-SSM}\left[c_1^{(\ell)}, c_2^{(\ell)}, \ldots, c_i^{(\ell)}\right] \tag{13}$$

The operations in Equations (12) and (13) correspond to Conv and SSM operations, respectively. The gating mechanism $g_i^{(\ell)}$ is given by:

$$g_i^{(\ell)} = \text{SiLU}\left[W_g^{(\ell)} h_i^{(\ell-1)}\right] \tag{14}$$

The formulas and concepts used here are adapted from Sharma et al. (2024).

Compared to Transformer, Mamba's design enables more efficient parallel training and effectively captures dependencies in sequences, making it suitable for various natural language processing tasks.

C.1.3 Vision and Multi-modal Models

In the realm of vision and multi-modal models, various architectures have emerged, each with its unique approach to tackling complex visual tasks. For example, **GANs** (Generative Adversarial Nets) (Goodfellow et al., 2014) consist of two neural networks: a generator and a discriminator. Through adversarial learning, the generator aims to produce realistic data samples (such as images), while the discriminator attempts to distinguish between real and generated data. **Diffusion Model** (Li et al., 2023c; Sohl-Dickstein et al., 2015) is a powerful tool for generating high-quality images and data. It simulates a diffusion process by gradually adding and removing noise to achieve data generation. **ResNet** (Residual Network) (He et al., 2016) introduced residual learning, revolutionizing deep network training by improving efficiency

and performance through skip connections. **ViT** (Vision Transformer) (Dosovitskiy et al., 2021) integrated the Transformer architecture into vision tasks, capturing long-range dependencies by processing image patches.

C.2 Knowledge Mechanisms in Other Architectures

Surprisingly, similar mechanisms as those found in transformer-based LLMs have also been discovered in other architectural models. Specifically, Mamba employs the knowledge memorization mechanism similar to Transformer (Sharma et al., 2024). Vision and multi-modal architectures also adopt function region (*Modular Region Hypothesis*) for knowledge utilization (Pan et al., 2023a; Schwettmann et al., 2023a; Koh et al., 2020; Bau et al., 2017), e.g., multi-modal neuron regions are responsible for multi-modal tasks. Besides, the *connections hypothesis* between neurons is found in vision architecture models (Olah et al., 2020). Olah et al. (2020) further suggest that different types of knowledge reuse partial components, e.g., cars and cats reuse the same neurons (*Reuse Hypothesis*). As for the Dynamic Intelligence Hypothesis, it inherently focuses on entire artificial neural models. Generally, neural models across various architectures, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces (Huh et al., 2024). These neural models may tend to share similar knowledge mechanisms and imagination (Zhou et al., 2024)

C.3 Machine and Human

Analogous to the Hominidae Family in biological taxonomy, artificial neural models can be regarded as Neural Model Family:

- Family: Neural Model, likened to "Hominidae".
- Genus: Transformer architecture, Mamba architecture, etc., likened to "Homo" and "Pan".
- Species: BERT, GPT, Llama, Mistral, Mamba, etc., likened to "Sapiens", "Pan troglodytes", and "Pan paniscus".

Metaphorically, Llama-7B, Llama-13B, Llama-70B, etc., can be viewed as the infancy, childhood, and adulthood of humans. Shah et al. (2024) further find that, regardless of model size, the developmental trajectories of PLMs consistently exhibit a window of maximal alignment with human cognitive development. Therefore, we hypothesize that **artificial neural networks (machine) and biological neural networks (human) tend to converge to universality intelligence.** In other words, huamn and machine share similar features and circuits.

Specifically, extensive evidences demonstrate that machine and human share the same mechanism of knowledge memorization, i.e., modular region and connection (de Schotten et al., 2022). The activations of modern language models can also linearly map onto the brain responses to speech (Caucheteux et al., 2023). Caucheteux et al. (2023) pioneer the explanation via predictive coding theory: while transformer-based LLMs are optimized to predict nearby words, the human brain would continuously predict a hierarchy of representations that spans multiple timescales. The above phenomenon indicates that machine and human share similar underlying mechanisms of knowledge (Sucholutsky et al., 2023; Chan et al., 2023; Kornblith et al., 2019), irrespective of their specific configurations, process and comprehend information. This could be due to inbuilt inductive biases (Sæbø and Brovold, 2024) in neural networks or natural abstractions (Chan et al., 2023) – concepts favored by the natural world that any cognitive system would naturally gravitate towards (Bereska and Gavves, 2024).

D Tools for Mechanism Analysis

Numerous tools exist for interpreting knowledge mechanisms in LLMs. *TransformerLens* (Nanda and Bloom, 2022) is a library for the mechanistic interpretability using observation and intervention. TransformerLens allows users to cache, remove, or replace internal activations during model running. *XMD* (Lee et al., 2023a) provides various forms of feedback via an intuitiveness, which enable explanations align with the user feedback. *NeuroX* (Dalvi et al., 2023) implements various interpretation methods under a unified API then provides interpretability of LLMs. *PatchScope* (Ghandeharioun et al., 2024) is a tool developed by Google that employs a novel model to elucidate the hidden states in the original model. *Transformer Debugger* (Mossing et al., 2024), an interpretability tool from OpenAI, utilizes GPT-4 and sparse auto-encoders to explain language neurons. Sparse autoencoders (Gao et al., 2024b) leverages sparse auto-encoders to extract interpretable fea-

tures from a language model by reconstructing activations from a sparse bottleneck layer. *Transcoders* (Dunefsky et al., 2024) decomposes model computations involving MLPs into interpretable circuits.

E Application of Knowledge Mechanism

The mechanism analysis of knowledge utilization and evolution may provide an avenue to construct more efficient and trustworthy models in practice.

E.1 Efficient LLMs

Researchers have been working to reduce the cost of training and inference for LLMs through various optimization strategies, including architecture (Ainslie et al., 2023; Fedus et al., 2022), data quality (Kaddour, 2023), parallelization (Qi et al., 2024), generalization theory (Zhang et al., 2024d), hardware (Dey et al., 2023), scaling laws (Hoffmann et al., 2022), optimizer (Liu et al., 2023a), etc. The underlying knowledge mechanisms offer LLMs new potential for efficiently storing, utilizing, and evolving knowledge.

For **knowledge storage and utilization** in LLMs, *knowledge (memory) circuit* provides the theory to decompose the knowledge computations of an LLM into smaller, recurring parts (Yang et al., 2024b). These smaller parts guide the determination of which types of knowledge should be encoded into parameters. Therefore, Memory^3 (Yang et al., 2024b) designs an explicit memory mechanism for Transformer-based LLMs, alleviating the burden of parameter size. Specifically, Memory^3 designs external information, explicit memory, and implicit memory for different usage frequencies, reducing writing and reading costs. For **knowledge evolution**, the knowledge mechanism analysis inspires *editing* and *model merging*. The details of editing technologies can be found in §3.2. Model merging technologies [9] leverage parameter directions to combine multiple task-specific models into a single multitask model without performing additional training rather than training from scratch. For instance, Task Arithmetic (Ilharco et al., 2023) identifies the weight directions of task capabilities in different models, and then integrates a more powerful model by arithmetic operations on weight directions. TIES (Yadav et al., 2023) resolves parameters directions conflicts, and merges only the parameters that are in alignment with the final agreed-upon sign. Akiba et al. (2024) further propose evolutionary optimization of model merging, which automatically discovers effective combinations of open-source models, harnessing their group intelligence without requiring extensive training data or computational resources. Besides, the Lottery Ticket Hypothesis (Frankle and Carbin, 2019) provides a cornerstone for *model compression*, generalizing across various datasets, optimizers, and model architectures (Morcos et al., 2019; Chen et al., 2021). However, model compression often limits the success of editing and model merging (Kolbeinsson et al., 2024). This phenomenon poses challenges for practical implementations, highlighting the need for more effective strategies.

[9] Model merging also includes methods that directly interpolation (Goddard et al., 2024) or randomly fusing (Yu et al., 2023b), ignoring parameter directions. These methods are naive and are not the focus of our discussion here.

E.2 Trustworthy LLMs

Numerous studies investigate the underlying causes of security risks (Reuel et al., 2024; Ren et al., 2024b; Li et al., 2024a; Bengio, 2024; Bengio et al., 2024; Dalrymple et al., 2024). In particular, Wei et al. (2023) delve into the safety of LLM and reveal that the success of jailbreak is mainly due to the distribution discrepancies between malicious attacks and training data. Geva et al. (2022b) and Wang et al. (2024b) further discover that some parameters within LLMs, called toxic regions, are intrinsically tied to the generation of toxic content. Ji et al. (2024) even conjecture that LLMs resist alignment. Therefore, traditional aligned methods, DPO (Rafailov et al., 2023) and SFT, seem to merely bypass toxic regions (Lee et al., 2024; Wang et al., 2024b), making them susceptible to other jailbreak attacks (Zhang et al., 2023d).

Inspired by the knowledge mechanism analysis in LLMs, a promising trustworthy strategy may be **designing architecture and training process** during the pre-training phase to encouraging modularity (Liu et al., 2024c, 2023b), sparsity (Chughtai et al., 2023), and monosemanticity (Bricken et al., 2023; Jermyn et al., 2022), which make the reverse engineering process more tractable (Jermyn et al., 2022; Bricken et al., 2023; Liu et al., 2024c; Tamkin et al., 2023). Yet, maintaining sparsity for a vast amount of world knowledge requires substantial resources, and whether monosemantic architecture can support advanced intelligence remains elusive. Besides, **machine unlearning** (Nguyen et al., 2022; Tian et al., 2024; Yao et al., 2023a) aims to forget privacy or toxic information learned

by LLMs. However, these unlearning methods suffer overfitting, forgetting something valuable due to the difficulty of disentangling verbatim memorization and general capabilities (Huang et al., 2024c; Blanco-Justicia et al., 2024). Another alternative technique is **knowledge editing**, precisely modifying LLMs using few instances during the post-training stage (Mazzia et al., 2023; Yao et al., 2023b; Wang et al., 2023d; Hase et al., 2024; Qian et al., 2024b). Extensive experiments demonstrate that knowledge editing has the potential to detoxify LLMs (Yan et al., 2024). Specifically, (Wu et al., 2023a) and Geva et al. (2022b) deactivate the neurons related to privacy information and toxic tokens, respectively. (Wang et al., 2024b) identify and then erases toxic regions in LLMs. However, knowledge editing also introduces side effects, such as the inability of the modified knowledge to generalize to multi-hop tasks (Zhong et al., 2023; Li et al., 2023d; Cohen et al., 2023; Kong et al., 2024) and the potential to impair the model's general capabilities (Gu et al., 2024; Qin et al., 2024). Therefore, recent efforts focus on **representation editing** instead of editing parameters in knowledge editing (Zou et al., 2023; Turner et al., 2023; Zhou et al., 2023b; Zhu et al., 2024). These representations (hidden states) within LLMs can trace and address a wide range of safety-relevant problems, including honesty, harmlessness, and power seeking. Later, (Wu et al., 2024) develop a family of representation finetuning methods to update new knowledge. (Zou et al., 2024) propose circuit-breaking (Li et al., 2023b), directly controlling the representations that are responsible for harmful outputs. However, these representation editing strategies require meticulous hyperparameter tuning for each task. More efficient optimization methods are needed to align with computational or temporal constraints.

F How to Explore More Knowledge from Interdisciplinary Inspiration?

How can LLMs continuously narrow the boundaries of dark knowledge and achieve higher level intelligence by leveraging the human experience of perpetual knowledge exploration throughout history? We may draw inspirations from the following interdisciplinary studies.

Neuroscience studies the structure and function of the brain at molecular, cellular, neural circuit, and neural network levels (Squire et al., 2012). Generally, both mechanism analysis in LLMs and neuroscience utilize observation and intervention methods to investigate the basic principles of knowledge learning and memory, decision-making, language, perception, and consciousness. The biological signals of the human brain and the internal activation signals in LLMs are capable of reciprocal transformation (Caucheteux et al., 2023; Feng et al., 2023a; Mossing et al., 2024; Flesher et al., 2021). Benefiting from advancements in neuroscience (Jamali et al., 2024; de Schotten et al., 2022; Lee et al., 2022a), mechanism analysis in LLMs has identified analogous function neurons and regions (Zhao et al., 2023a), and knowledge circuits (Yao et al., 2024). Besides, leveraging plasticity theory in neuroscience, LLMs explain the underlying technical support for intelligence (Sæbø and Brovold, 2024). In the future, mechanism analysis of LLMs may draw inspirations from neuroscience, guiding the next generation of artificial intelligence in organizing neural frameworks and in the storage and utilization of knowledge (Ren and Xia, 2024; Momeni et al., 2024; Yang et al., 2024b).

Cognitive Science focuses on the mind and its processes (Kolak et al., 2006; Baronchelli et al., 2013), which include language, perception, memory, attention, reasoning, emotion and mental state. Although cognitive science and neuroscience overlap in their research content, cognitive science focuses more on abstract knowledge such as mental states and emotions rather than specific knowledge. Therefore, Zhu et al. (2024) track beliefs of self and others (formulated as "Theory of Mind") in LLMs from the psychological perspective within cognitive science. (Wang et al., 2022) further observe social-cognitive skill in multi-agent communication and cooperation. Generally, there is potential to explore advanced cognitive capabilities in LLMs from the perspective of cognitive science (Vilas et al., 2024).

Psychology is the scientific study of mind and behavior, which include both conscious and unconscious phenomena, and mental processes such as thoughts, feelings, and motives. Benefiting from decades of research in human psychology, machine psychology aims to uncover mechanisms of decision-making and reasoning in LLMs by treating them as participants in psychological experiments (Hagendorff, 2023). Machine psychology may delve into mysteries of social situations and interactions shaping machine behavior, attitudes, and beliefs (Park et al., 2023a). Besides, group

psychology paves an auspicious path for exploring dynamics such as debates and collaboration among LLMs (agents). For instance, *Dunning–Kruger effect* (Mahmoodi et al., 2013; Brown and Esterle, 2020) in cognitive psychology filed describes that individuals with limited competence in a particular domain overestimate their abilities, and vice versa. This phenomenon may guide the final vote in group debates and discussions. Promisingly, psychology of learning can be applied to study prompt designs, boost learning efficiency, improve communication strategies, and develop feedback mechanisms for LLMs (Leon, 2024).

Education is the transmission of knowledge, skills, and character traits and manifests in various forms. Inspired by education in humans, Zhang et al. (2024a) categorize knowledge acquisition in LLMs into three distinct phases: recognition, association, and mastery. Besides, education instructs humans managing various types of conflicts: identifying inconsistencies in external information (inter-context conflict), deciding between external sources and internal memory (context-memory conflict), resolving memory confusion (internal memory conflict), and addressing cultural conflicts. The above knowledge conflicts and integration also exist in knowledge evolution of LLMs across individuals and groups (Dan et al., 2023). Fortunately, education facilitates humans in learning to learn. Can LLMs similarly self-evolve to continuously adapt to societal changes and requirements?

✨ **Remarks**: LLMs may improve their architecture and mechanisms for knowledge learning, storage, and expression, drawing inspiration from neuroscience. Besides, cognitive science and psychology provide promising alternatives for sophisticated intelligence, emergent capabilities and behaviors in evolution. Educational studies can inspire the learning strategy of LLMs, navigating conflicts and integrating knowledge during their evolution.

G Future Directions

G.1 Parametric VS. Non-Parametric Knowledge

LLMs can be conceptualized as parametric knowledge stores, where the parameters of the model—typically the weights of the neural network—encode a representation of the world's knowledge. This parametric approach to knowledge storage means that the knowledge is implicitly embedded within the model's architecture, and it can be retrieved and manipulated through the computational processes of the neural network (Allen-Zhu and Li, 2023b). In contrast, non-parametric knowledge storage involves methods where the knowledge is explicitly represented and can be directly accessed. Examples of non-parametric knowledge storage include knowledge graphs, databases, and symbolic reasoning systems, where knowledge is represented as discrete symbols or facts. Parametric knowledge enables LLMs to deeply compress and integrate information (Huang et al., 2024d; Shwartz-Ziv and LeCun, 2024), allowing them to generalize and apply this knowledge across various contexts. This is akin to LLMs mastering the mathematical operation rule of "mod" through parametric knowledge, enabling them to generalize and seamlessly solve all mod-related problems (Pearce et al., 2023; Hu et al., 2024). Conversely, non-parametric knowledge requires extensive searches across the knowledge space for each user query. Subsequently, Wang et al. (2024a) also prove that non-parametric knowledge severely fails in complex reasoning tasks, with accuracy levels approaching random guessing. Unfortunately, parametric knowledge within LLMs is opaque, often encountering challenges such as interpretability issues, outdated information, hallucinations, and security concerns.

Addressing these issues often requires leveraging external non-parametric knowledge, which offers transparency, flexibility, adaptability, and ease of operation. However, *augmenting parametric knowledge* in LLMs with non-parametric knowledge (Yang et al., 2024b; Luo et al., 2023; Wen et al., 2023; Ko et al., 2024) remains an ongoing challenge due to retrieval accuracy from haystack, context lengths, and resources [10] limitations (Shang et al., 2024; Zhao et al., 2024b). Besides, simultaneously retrieving relevant information from a long context and conducting reasoning is nearly impossible in reasoning-in-a-haystack experiments (Shang et al., 2024). Similarly, *augmenting non-parametric knowledge*—either by distilling knowledge from an LLM's parametric knowledge (West et al., 2022; Kazemi et al., 2023) or by using it to parse text directly (Zhang et al., 2023b)—also poses significant challenges. Moreover, Yang et al. (2024b) propose *a novel explicit memory that lies*

[10] On the one hand, storing large amounts of non-parametric knowledge requires a lot of space and high maintenance costs. On the other hand, retrieving information for each user query is very resource-intensive.

between parametric and non-parametric knowledge. LLM with explicit memory enjoys a smaller parameter size and lower resource consumption for retrieving external non-parametric knowledge.

Generally, **inspired by the knowledge mechanisms analysis in LLMs, we have the potential to develop more architectural and learning strategies for organizing knowledge within LLMs**. These efficient LLMs (Sastry et al., 2024) are advancing toward lower GPU, computation, and storage resource requirements, as well as smaller model sizes by combining the strengths of parametric and non-parametric knowledge (Yang et al., 2024b; Momeni et al., 2024; Chen, 2024; Pan et al., 2024b, 2023b).

G.2 Embodied Intelligence

The current LLM still cannot be regarded as a truly intelligent creature (Bender and Koller, 2020; Bisk et al., 2020). The process of human language acquisition is not merely a passive process of listening to language. Instead, it is an active and interactive process that involves engagement with the physical world and communication with other people. To enhance the current LLM's capabilities and transform it into a powerful agent, it is necessary to enable it to learn from multimodal information and interact with the environment and humans.

Multimodal LLMs. The integration of multiple modalities is a critical challenge in the field of LLMs and embodied AI. While LLMs have demonstrated impressive capabilities when processing language data, their ability to seamlessly incorporate and synthesize information from other modalities such as images, speech, and video is still an area of active research. However, the current multi-modal model faces challenges, particularly in complex reasoning tasks that require understanding and integrating information from both text and images.

Recent studies (Huang et al., 2024a; Chen et al., 2024b) have highlighted the discrepancy between the model's performance in language tasks and its ability to integrate knowledge from different modalities effectively. These findings suggest that current models often prioritize linguistic information, failing to fully exploit the synergistic potential of multimodal data (Wang et al., 2024c). There are some pioneering efforts in this direction (Pan et al., 2024a; Schwettmann et al., 2023a) , aiming to uncover the mechanisms by which multi-modal models store and retrieve information. Despite these advancements, there is still a need for further exploration to deepen our understanding of multi-modal knowledge storage.

Self-evolution. As discussed in the previous part, current language models are mainly based on tuning to gain knowledge, which requires a lot of training and high-quality data. These learnings are passive whereas, to be a human, evolution usually also undergoes communication and interaction. As an intelligent agent, the models should be able to learn through interactions and learn by themselves spontaneously. Recently, some work has attempted to enable the model to learn by themselves (Zhang et al., 2024b) or learn by interaction with the environment (Xu et al., 2024a; Xi et al., 2024). By integrating self-evolving mechanisms, models can continuously update their knowledge base and improve their understanding without relying solely on manually curated datasets. This not only reduces the dependency on large-scale labeled data but also allows the models to adapt to evolving linguistic norms and cultural contexts over time.

G.3 Domain LLMs

The success of general-purpose LLMs has indeed inspired the development of domain-specific models that are tailored to particular areas of knowledge (Calderon and Reichart, 2024), such as biomedicine (Yu et al., 2024a; Moutakanni et al., 2024), finance (Yang et al., 2023), geoscience (Deng et al., 2023), ocean science (Bi et al., 2024), etc. However, unlike human language, the knowledge of these different domains bears specific characteristics. It remains unclear whether LLMs can acquire complex scientific knowledge or if such knowledge still resides within the realm of current dark knowledge. Furthermore, does domain-specific knowledge such as mathematics share the same underlying mechanisms as textual knowledge (Bengio and Malkin, 2024), or does it exhibit more intricate mechanisms of knowledge acquisition? Currently, there is a relative lack of research focusing on the mechanism of these domain-specific knowledge and there is an increasing recognition of the importance of developing a deeper understanding of these mechanisms.

Data sparsity and diversity in domain-specific models pose another challenge. Sparsity is usually caused by confidentiality, privacy, and the cost of acquisition in specialized fields. As for diversity, the presentation of knowledge varies across

different fields. For instance, in the biomedical domain, knowledge includes complex biological concepts such as the structure and function of proteins and molecules. This requires models to integrate understanding that extends beyond natural language, often involving graphical representations like chemical structures, which cannot be directly expressed in text. Similarly, in fields such as finance and law (Lai et al., 2023), models must engage in sophisticated reasoning and decision-making processes based on domain-specific knowledge. Hence, the critical tasks of collecting high-quality data for domain-specific models (including synthetic data generation) and effectively embedding domain knowledge into LLMs require immediate attention.