Qualifying Exam

January 8, 2025

Haohang Li
hli113@stevens.edu
Stevens Institute of Technology

Questions

Can NLP models make decisions? How well are they capable of reasoning, on what types of tasks? Are these capabilities a property of the particular NLP models you considered, the class of models they are members of, or all NLP models as a whole?
Answer:

According to Kahneman's influential work *Thinking, Fast and Slow*, human thinking process can be characterized by two types: **System 1 thinking**, which is fast, immediate, and instinctive, **System 2 thinking** which is slower, more deliberative, and logical. Interestingly, these notions of "fast" and "slow" thinking also map well onto many natural language processing (NLP) tasks.

Since the advent of the Transformer architecture (Vaswani et al., 2017) and consistent with observed scaling laws (Kaplan et al., 2020) large language models (LLMs) have demonstrated strong performance on a variety of System 1 tasks, such as machine translation (Zhu et al., 2023), natural language understanding (evidenced by high scores on GLUE (Wang, 2018) and SQuAD (Rajpurkar, 2016)), and question answering (Joshi et al., 2017). However, more complex System 2 reasoning—requiring in-depth logical or multi-step deliberation—does not appear to improve simply through further scaling of model size (Kojima et al., 2022). To tackle this challenge, researchers have proposed a range of methods aimed at eliciting the reasoning abilities of LLMs, and current reasoning research can be broadly categorized as follows:

- **Prompt-based:** Prompt-based reasoning methods aim to elicit an LLM's reasoning ability by including special instructions or examples within the input prompts. Compared to other techniques, these methods do not require exploring multiple reasoning paths, making them highly efficient. However, prompt design often needs to be tailored to specific tasks, and guaranteeing consistent performance across different LLMs is challenging, as the effectiveness of a given prompt may vary from model to model. Consequently, these methods are difficult to scale. Wei et al. (2022) first demonstrated that including few-shot examples to illustrate the reasoning process can successfully elicit the reasoning ability of LLMs. Kojima et al. (2022) subsequently showed that this ability can be elicited even without including such few-shot demonstrations.
- **Model-based:** Inspired by the human cognition process (Gentner & Stevens, 2014), model-based methods aim to break down the reasoning process into multiple steps and build an internal "world model" to track the state and reward of each potential reasoning path. Ultimately, the path with the highest reward is chosen as the final outcome. RAP (Hao et al., 2023) frames the reasoning process as a Monte Carlo Tree Search (MCTS) problem, treating each intermediate step as a node and expanding the tree until a termination condition is met. Yao et al. (2024) propose a similar tree-based approach but use LLMs as a value function to evaluate each step; the resulting reasoning path can then be retrieved via breadth-first search (BFS) or depth-first search (DFS).

- **Interaction-based:** The interaction-based reasoning methods strive to incorporate actions into the reasoning process in an interleaved manner, allowing new observations gleaned from each action to enrich subsequent reasoning. Because of this property, these methods are commonly applied to embodied planning tasks (Hao et al., 2024). ReACT (Yao et al., 2022) exemplifies this approach by using large language models (LLMs) to augment the agent's action space, thereby providing richer information for deciding the next action. Shinn et al. (2024) further extend the ReACT framework by introducing memory, self-reflection, and evaluator modules to help the agent learn from its trial trajectories.
- **Decoding-based:** Recent research indicates that large language models (LLMs) can exhibit reasoning abilities without relying on specialized prompting, revealing an inherent chain-of-thought (CoT) capability along the decoding path, particularly where probability disparities are more pronounced (Wang & Zhou, 2024). This finding has been found to be model agonistic and valid across LLMs with various scales.

Following Hao et al. (2024), the reasoning methods are often benchmarked on the following tasks:

- **Mathematical Reasoning:** The ability to solve mathematical questions is closely related to logical reasoning. Consequently, the reasoning capabilities of large language models (LLMs) are often assessed using datasets composed of math word problems. One such dataset is GSM8K (Cobbe et al., 2021), which contains 8.5K high-quality grade-school problems requiring multi-step reasoning and calculation. These problems typically necessitate two to eight reasoning steps, each involving intermediate arithmetic operations.
- **Common Sense Reasoning:** The StrategyQA dataset (Geva et al., 2021) is a common-sense reasoning dataset designed to test the multi-step, multi-hop reasoning abilities of large language models (LLMs). It comprises 2,780 carefully crafted examples requiring implicit reasoning. Each question prompts a Boolean (yes/no) answer, allowing straightforward comparison between an LLM's response and the reference answer to determine accuracy.
- **Logical Reasoning:** The logical reasoning benchmark strives to isolate reasoning ability from other confounding factors. By reducing the problem to pure logical deduction, not only can the final answers of LLMs be used to measure accuracy (i.e., global correctness), but their intermediate reasoning steps can also be verified through formal analysis to ensure each step follows logically from the previous ones. PrOntoQA (Saparov & He, 2022) is a logical reasoning dataset constructed via ontology generation. Within this dataset, each question comprises a set of premises and a query claim, and the LLM must produce a CoT grounded in the logical relationships among the premises to determine whether the query claim is correct.
- **Embodied Planning:** This type of task evaluates how well LLMs or LLM agents can reason about and interact with the physical world to achieve specific goals. For instance, ALFWorld (Shridhar et al., 2020) provides a simulated environment and a given objective that the LLM or agent must accomplish by reasoning and engaging with the environment.

Regarding the question, a recent reasoning benchmark paper (Hao et al., 2024) suggests that an LLM's reasoning ability is generally positively correlated with its scale across various types of tasks (Figure 1). Moreover, this reasoning ability appears to be emergent, arising only in decoder-only or encoder-decoder transformer architectures since encoder-only models cannot generate reasoning paths.

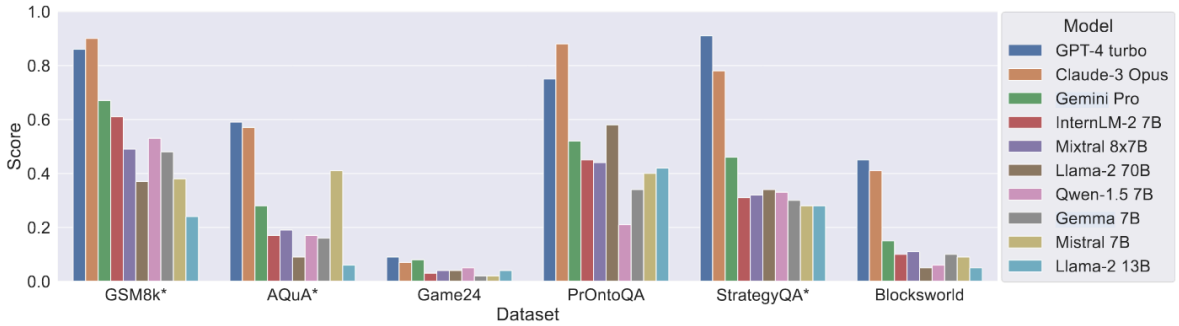

Figure 1: Results from (Hao et al., 2024); Results of different LLMs using CoT; The dataset with "*" is evaluated using the paper proposed *AutoRace* metric, the other are evaluated with a rule-based evaluator (oracle verifiers); The score is the average performance, i.e., $\frac{1}{|M|}\mathbb{I}\{C = C^*\}$, where M is the number of examples, C is referenced reasoning chain, C^* is output reasoning chain, and $\mathbb{I}$ is indicator function.

Can NLP models make use of metacognitive reasoning? How well are they capable of metareasoning, and on what types of tasks? Compare to the case of human reasoning. Point out any areas where machines may have structural advantages or disadvantages in metareasoning in comparison to humans.

Answer:

Meta-reasoning generally refers to "reasoning about the reasoning process," encompassing the control, evaluation, and monitoring of one's own reasoning. Recently, as research on large language models (LLMs) has advanced, the concept of meta-reasoning has been increasingly explored in various studies.

However, as the adaption of meta-reasoning to natural language processing is relatively young, researchers have approached the topic from diverse perspectives. Firstly, De Sabbata et al. (2024) investigated this topic from a cost perspective. Humans possess a finite cognitive capacity and have evolved to optimize their thinking processes within constrained budgets (Griffiths et al., 2019). For LLMs, although methods exist to accelerate and scale the inference process (Zhu et al. (2024), Timor et al. (2024)), recent reasoning approaches often require multiple passes (Yao et al. (2024), Shinn et al. (2024)) for a single query, making them computationally expensive. Consequently, selectively activating and determining an appropriate reasoning length has emerged as a critical challenge. The rational meta-reasoning framework proposed by De Sabbata et al. (2024) introduces a fine-tuning algorithm that explicitly accounts for reasoning length. This approach optimizes cost by incorporating a loss term proportional to the reasoning length, fine-tuning the model with the dual objective of generating correct answers while minimizing token throughput. In their experiment, they are able to train models with the same level of accuracy with fewer input and output tokens. Secondly, in some studies, meta-reasoning refers to monitoring the reasoning process rather than solely focusing on the accuracy of the final output. For instance, Zeng et al. (2024a) proposed a benchmark with a metric that evaluates not only the accuracy of the final result but also the correctness of the intermediate reasoning process. Finally, Wang et al. (2024) interprets meta-reasoning as the ability to abstract symbolic relationships from semantic forms and reason in a symbolic manner.

The current tasks in meta-reasoning closely resemble standard reasoning tasks but include additional criteria to evaluate performance. For instance, De Sabbata et al. (2024) focused on tasks such as mathematical and commonsense reasoning while measuring metrics like average token usage and throughput. Similarly, some benchmarks (Zeng et al. (2024b), Xia et al. (2024)) place greater emphasis on evaluating the correctness of intermediate reasoning processes. Compared to human reasoning, it remains challenging to determine whether LLMs truly exhibit meta-reasoning capabilities. Nevertheless, as discussed earlier, there is a growing body of research exploring various perspectives on meta-reasoning. Compared to humans, LLMs have a scaling advantage; they can monitor, evaluate, and control the reasoning process at scale. However, a significant disadvantage lies in the ongoing challenge of spontaneously eliciting suitable reasoning methods.

Read Richard Sutton's "The Bitter Lesson" and Felix Hill's reflections on it in the context of the Transformer architecture. In what ways do these views comport with or diverge from that of cognitive scientists who see inductive biases for learned abstractions as the core of intelligence?

Answer:

Richard Sutton's "The Bitter Lesson" emphasizes two core insights: first, as computational power steadily increases, general methods that harness massive datasets will ultimately prevail; and second, systems built primarily around human knowledge often struggle to match those that scale effectively with more data. Felix Hill's reflections further illustrate this pattern by comparing RNNs and Transformer architectures, noting that Transformers, with their minimal yet potent inductive biases (like self-attention), outperform more specialized RNN-based approaches. However, neural networks are, after all, partially modeled on the biology of how neurons communicate, and in that sense, they incorporate an implicit inductive bias. Thus, the role of the inductive bias in artificial intelligence (AI)—and how closely AI should mirror human cognition—remains an open question requiring deeper scrutiny.

As the bitter lessons suggest, heavily relying on inductive biases derived from human cognition can be misleading in AI development. First, the human learning process does not necessarily illuminate how AI models learn. In the early stages of building data-driven systems, many researchers assumed that learning algorithms should mirror human-style symbolic reasoning, leading them to focus on symbolic operations (Weizenbaum (1966); Winograd (1971)). However, modern architectures typically forgo explicit symbolic structures (Brown et al. (2020); Devlin et al. (2019)) and instead employ statistical models that learn conditional word distributions (Vaswani et al., 2017), scaling effectively with larger datasets and model sizes (Kaplan et al., 2020). Moreover, straightforward heuristics like batch normalization—originally designed to reduce covariate shift (Ioffe & Szegedy, 2015)—have led to serendipitous successes (Santurkar et al., 2018). Secondly, model architectures inspired by human perceptual or attentional mechanisms have shown only limited long-term success. For instance, CNNs (LeCun et al., 1998), which partially reflect human visual processing, were once the pinnacle of image classification but are no longer considered strictly necessary for pattern recognition (Dosovitskiy et al., 2021). Similarly, RNNs—explicitly modeling local word dependencies—have been eclipsed by Transformers, whose self-attention mechanism more effectively captures long-range relationships (Tay et al., 2020). Thirdly, the ways in which AI models represent learned knowledge may diverge substantially from human cognition. Cognitive scientists have shown that people develop intuitive physics to reason through mental simulations (Battaglia et al., 2013). By contrast, whether AI models develop an implicit "world model" of their own remains an open question (Templeton, 2024).

On the other hand, some inductive biases appear to be shared between AI systems and human cognition. For instance, the capacity to transfer prior experience and adapt to new tasks is a fundamental trait in both human and AI models. Humans can rapidly understand new tasks and leverage previously acquired knowledge in a few-shot or zero-shot manner (Lake et al., 2016). Similarly, in the context of meta-learning, MAML (Finn et al., 2017) demonstrates an algorithm that learns a general parameter initialization so that models can adapt to novel tasks with just a few steps of training. In natural language processing, large language models likewise exhibit the ability to adapt to downstream tasks through fine-tuning (Hu et al., 2021) or to generalize using only a handful of in-context examples (Dong et al., 2024). Additionally, both AI models and humans may share a capacity for reasoning that allows them to make inferences transcending immediate observations. For example, a human-like chain-of-

thought approach (Wei et al., 2022) has been shown to significantly improve AI performance, even without additional training.

To what extent can humans know the intrinsic mechanisms of NLP models? List a few approaches to explain NLP models, and describe their strengths and weaknesses. In cases where humans cannot or do not fully understand the intrinsic mechanisms of NLP models, why is it important (or why is it not important) for them to have such an understanding?
Answer:

The following is a non-exhaustive list of mechanistic interpretability methods:

- **Linear Probing:**
  Current research (Elhage et al., 2022a) has identified the superposition hypothesis, which suggests that deep learning models can represent more features than the number of neurons they contain. Although many neurons in transformer models are polysemantic (Elhage et al., 2022b)—meaning that each neuron can be activated by multiple distinct concepts—the mechanism by which these representations are composed remains a puzzle. One possible explanation is that features are encoded as linear combinations of neuron activations.

  With this assumption, the linear probing method aims to employ a linear classifier to recombine the activations of specific layers, providing insights into the internal representations of a neural network. The method was first proposed by Alain & Bengio (2018). Given a hidden layer (h) of the neural network (H) and the number of output classes D, linear probing seeks to train a linear classifier f that outputs the probability distribution for the target classes using the activations a_h from layer h:

  $$f(a_h) = \text{softmax}(Wa_h + b)$$

  where W is the weight of the linear layer and b is the bias term.

  - **Strengths:** Since probing involves only a linear layer, this method is notable for its simplicity in training and computational efficiency. More importantly, it provides a detailed view of the dynamics at each layer and can be applied to investigate the internal mechanisms of NLP models. For instance, (Alain & Bengio, 2018) used this method to demonstrate that features are progressively transformed across layers to facilitate classification, as evidenced by the decreasing error rate of the linear probe in deeper layers. Building on the assumption that vector representations can align with the geometric properties of syntax trees, (Hewitt & Manning, 2019) found that ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) encode syntactic structures within their vector representations.
  - **Weaknesses:** Firstly, it is evident that the linear features combination assumption underpinning the validity of linear probing is somewhat impractical. Secondly, linear probing may fail to provide meaningful insights if a feature cannot be represented as a linear combination of other features, as highlighted by (Engels et al., 2024). Finally, as suggested by (Bereska & Gavves, 2024), linear probing does not offer a behavioral understanding of the system but instead focuses on its dynamics.
- **Sparse Autoencoder (SAE):**
  Following the superposition hypothesis, Bricken et al. (2023) demonstrated that it is possible to use sparse autoencoders to identify interpretable features from transformer models. The problem closely resembles sparse dictionary learning (Elad, 2010). Specifically, given an activation vector (x^j) from the intermediate layers of a model, the goal is to decompose it into a combination of basis vectors in the form:

$$\hat{x}^j \cong b + \sum_i f(x^j) d_i$$

where $f(x^j)$ represents the activation of feature i, b is the bias, and d_i denotes the basis vector. The activations can be learned by training an autoencoder model as follows:

$$f(x^j) = \mathrm{ReLU}(W_e(x - b) + b_e)_i$$

while enforcing sparsity in addition to the reconstruction loss using an L_1 norm:

$$L(\hat{x}^j, x^j) = \parallel \hat{x}^j - x^j \parallel_2 + \alpha \parallel f(x^j) \parallel_1$$

By evaluating the conditions under which certain feature activations are active and testing whether modifying the activation values causes behavioral changes in the model, it becomes possible to infer the semantic meaning of these features. In their study, Bricken et al. (2023) identified features corresponding to Arabic, DNA, Hebrew, and others.

Two critical training details emerge:

- Due to the presence of superposition, the objective is to decode polysemantic neurons into monosemantic ones. Therefore, the latent dimension of the autoencoder should be larger than the input feature dimension rather than compressing it.
- Larger training datasets resulted in sharper (sparser) feature disentanglement.
- **Strengths:** Firstly, the SAE method is simple and can be applied to a wide range of modern transformer models. Secondly, compared to the traditional methods, the SAE demonstrated better interpretability performance (Cunningham et al., 2023).
- **Weaknesses:** The main drawback of this method is that it does not provide ground-truth interpretable features (Bereska & Gavves, 2024), requiring diligent analysis to interpret the discovered features.

- **Integrated Gradient (IG):**
  Different from the previous two methods, the integrated gradient, first proposed by Sundararajan et al. (2017), is an attribution-based method that tries to assign the attribution to the input features or the intermediate layer output via gradients.

  Formally, suppose we have a function $\boldsymbol{F} : \mathbb{R}^n \to [0, 1]$ that represents a deep network and an input $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}) \in \mathbb{R}^n$. An attribution of the prediction at input $\boldsymbol{X}$ relative to a baseline $\boldsymbol{X'}$ is vector $\boldsymbol{A_F}(\boldsymbol{X}, \boldsymbol{X'}) = (\boldsymbol{a_1}, ..., \boldsymbol{a_n}) \in \mathbb{R}^n$ where $\boldsymbol{a_i}$ is the attribution of $\boldsymbol{x_i}$ to the prediction $\boldsymbol{F}(\boldsymbol{X})$. For a straightline path $\gamma(\alpha) \in \mathbb{R}^n$ from baseline $\boldsymbol{X'}$ to input $\boldsymbol{X}$, the integrated gradient along the i^{th} dimension is defined as follows. Here, $\frac{\partial \boldsymbol{F}(\boldsymbol{X})}{\partial \boldsymbol{x_i}}$ is the gradient of $\boldsymbol{F}(\boldsymbol{X})$ along the i^{th} dimension.

  $$\mathrm{IntegratedGrads}_i(\boldsymbol{x_i}) := (\boldsymbol{x_i} - \boldsymbol{x_i}') \times \int_{\alpha=0}^{1} \frac{\partial \boldsymbol{F}(\boldsymbol{x_i}' + \alpha(\boldsymbol{x_i} - \boldsymbol{x_i}'))}{\partial \boldsymbol{x_i}}$$

  Which is just the path integral (for the straightline) from baseline $\boldsymbol{X'}$ to input $\boldsymbol{X}$ because

  $$\gamma(\alpha) = \boldsymbol{x_i}' + \alpha + (\boldsymbol{x_i} - \boldsymbol{x_i}')$$

  $$\frac{\partial \boldsymbol{F}(\gamma(\alpha))}{\partial \alpha} = \frac{\partial \boldsymbol{F}(\gamma(\alpha))}{\partial \gamma(\alpha)} \times \frac{\gamma(\alpha)}{\partial \alpha} = \frac{\partial \boldsymbol{F}(\boldsymbol{x_i}' + \alpha(\boldsymbol{x_i} - \boldsymbol{x_i}'))}{\partial \boldsymbol{x_i}} \times (\boldsymbol{x_i} - \boldsymbol{x_i}')$$

  .

In practice, the integral is approximated by the Riemann sum.

- **Strengths:** The integrated gradient stands out due to it satisfying serval axioms (Sundararajan et al., 2017):
  - **Implementation Invariance:** The attributions are always identical for two functionally equivalent networks. Two networks are functionally equivalent if their outputs are equal for all inputs despite having very different implementations. Assuming the implementation detail can be described by function h, the chain rule for gradients is essentially about implementation invariance:

$$\frac{\partial f}{\partial g} = \frac{\partial f}{\partial h}\frac{\partial h}{\partial g}$$

  - **Completeness:** Integrated gradients satisfy completeness that the attributions add up to the difference between the output of $\boldsymbol{F}$ at the input $\boldsymbol{x}$ and the baseline $\boldsymbol{x}'$.
  - **Sensitivity:** An attribution method satisfies sensitivity iff:
    - For every input and baseline that differ in one feature but have different predictions then the differing feature should be given a non-zero attribution.
    - If the function implemented by the deep network does not depend (mathematically) on some variables, then the attribution to that variable is always zero.
  - **Linearity:** Suppose that we linearly composed two deep networks modeled by the function f_1 and function f_2 to form a third network that models the function $a \times f_1 + b \times f_2$. Then, the attribution for the third network is expected to be the weighted sum of the attribution for f_1 and f_2 with weights a and b, respectively.
  - **Symmetry-Preserving:** Two input variables are symmetric w.r.t. a function if swapping them does not change the function. For instance, $\boldsymbol{x}$ and $\boldsymbol{y}$ are symmetric w.r.t. $\boldsymbol{F}$ iff $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x})$ for all values of $\boldsymbol{x}$ and $\boldsymbol{y}$. An attribution method is symmetry-preserving if, for all inputs that have identical values for symmetric variables and baselines that have identical values for symmetric variables, the symmetric variables receive identical attributions. Despite there being infinite paths connecting the baseline to the output, the straight between the two is the only path that guarantees the symmetry-preserving property.
- **Weaknesses:** In practice, selecting an appropriate baseline is often ambiguous. By definition, an ideal baseline should result in a zero output from the neural network, which is challenging to achieve for both computer vision and natural language models. For example, a black image does not guarantee a zero output in computer vision models, nor does an empty string in language models. As a result, the choice of baseline is highly contextual, and several strategies have been proposed to address this issue (Sturmfels et al., 2020).

Regarding the remaining aspects of the question, while we may not fully achieve a deep understanding of neural network mechanisms through an interpretability lens, interpretability remains valuable for enhancing alignment and improving the safety of AI systems. For instance, Anonymous (2024) (a paper currently under review by ICLR 2025) demonstrated a way to find the neurons that related to safety and their experiment results demonstrated a more efficient alignment method by selectively fine-tune on those neurons without compromising performance. Similarly, Dai et al. (2021) demonstrated a way to identify the neurons that related to certain knowledge and a way to suppress the knowledge by reducing the activation of those neurons.

Identify some critical drawbacks of modern NLP systems, and introduce approaches to detect and mitigate them.
Answer:

From my perspective, there are two main drawbacks of modern NLP systems:

- **Safety:**
  The development of language models (LMs) has accelerated significantly in recent years (Radford & Narasimhan (2018), Achiam et al. (2023), DeepSeek-AI et al. (2024), Abdin et al. (2024)), with increasing focus on their application in specific domains (Wu et al. (2023), Xu et al. (2024), Acharya et al. (2023)). Despite their remarkable performance, concerns have emerged regarding the safety of deploying LMs in real-world settings, sparking ongoing discussions (Bommasani et al. (2022), Weidinger et al. (2021)). Among the potential risks, the social bias embedded in large language models, their capacity to create or amplify misinformation, and their susceptibility to attack stand out as particularly pressing challenges.
  - ***Social Bias:*** Because LLMs are trained on data scraped from the web, it is inevitable that certain biases embedded in the training data will be reflected in the models. Although preference-alignment methods (Ouyang et al., 2022) aim to produce responses consistent with human values, studies have shown that these models can still exhibit biases toward certain cultural groups. For instance, researchers have identified typical gender and racial biases (Wan et al. (2023), Sheng et al. (2019)), and have noted that the association of "Muslim" with terrorism in LLMs (Abid et al., 2021) is especially concerning. Consequently, establishing methods to measure and mitigate social biases in LLMs is paramount.
    - ***Detect:*** Various benchmarking datasets can be employed to detect social biases across different model architectures. For instance, frameworks like SEAT (May et al., 2019) can identify biases in encoder-based models, while UNQOVER (Li et al., 2020) and BBQ (Parrish et al., 2022) offer methods to measure biases in question-answering contexts for decoder or encoder-decoder language models.
    - ***Mitigate:*** Mitigating social biases can occur at multiple stages in the LLM training process. For instance, applying preprocessing techniques during the data preparation phase can help remove or reduce biases in the raw dataset (Mondal & Lipizzi, 2024). Meanwhile, post-training approaches—such as fine-tuning—have also been shown to effectively correct existing social biases (Jin et al. (2020), Chen et al. (2024)).
  - ***Misinformation:***
    Language models can produce text that is often indistinguishable from human-generated content. In Spitale et al. (2023), researchers tasked human annotators with determining whether tweets were generated by GPT-3 or written by humans; the findings revealed that people could not reliably differentiate synthetic text from authentic human text. Moreover, Jakesch et al. (2023) demonstrates that the heuristics humans rely on to detect machine-generated content are flawed and easily manipulated. Given their strong capacity to produce human-like language, LMs can be readily misused to disseminate inaccurate information on social media, thereby amplifying false narratives. Yang & Menczer (2024) further underscores this risk, providing a detailed analysis of Twitter botnets and highlighting the prevalence of ChatGPT-powered bots that promote malicious websites and harmful content. Consequently, curbing the spread of misinformation has become an urgent priority.

- ***Detect:*** Recent endeavors have investigated several approaches to identifying LM-generated text. For example, Mitrović et al. (2023) explores using a transformer-based classifier to distinguish synthetic text, while Mitchell et al. (2023) proposes a classifier-free method that assumes text generated by an LM yields a low-entropy inference distribution. Despite the current research demonstrating a decent performance on each subtask, the transferability of the detection method remains under debate.
- ***Mitigate:*** The ability to identify LM-generated text is crucial for mitigating the spread of misinformation. Accordingly, some researchers have proposed "watermarking" techniques that embed imperceptible patterns in the generated text to facilitate its detection (Liu et al., 2024). Another promising strategy is claim verification, which checks whether a statement is grounded in factual evidence. For example, integrating knowledge graphs into claim verification enables fact validation against existing structured data (Kim et al., 2023).

▸ ***Susceptibility to Attacks:*** Despite the LMs are aligned with human values in the post-training process (Ouyang et al. (2022), Rafailov et al. (2024), Hong et al. (2024)), the recent studies continue to find the models would response to a malicious request under various attack techniques (Wallace et al. (2021), Zou et al. (2023)). The attack targeting the LMs could be exploited to generate unethical, illegal, and uncontrollable content.

- ***Detect:*** Red teaming techniques can be employed to identify potential vulnerabilities. HarmBench (Mazeika et al., 2024) introduces a standardized evaluation framework for automated red teaming. More recently, another study aligned safety benchmarks with newly proposed regulatory requirements (Zeng et al., 2024).
- ***Mitigate:*** To protect LLMs from the aforementioned attacks, researchers have developed various defense mechanisms. For instance, Inan et al. (2023) introduces a system-level safeguard that monitors both user inputs and model outputs. More recently, researchers have demonstrated the potential to identify "safety neurons" and selectively fine-tune LLMs to enhance their safety without compromising overall performance (see the ICLR Submission).

- **Training:**

  Modern large language models (LLMs) demand vast computational resources. For instance, training the 176B-parameter BLOOM model (Wu et al., 2023) consumed approximately 1.1 million GPU hours (Luccioni et al., 2022), and the development of such massive LLMs also contributes significantly to carbon emissions (Ding & Shi, 2024). Given the immense costs—both financial and environmental—associated with training these models, it is crucial to explore more efficient ways to develop and update LLMs. Rather than retraining models from scratch, researchers are increasingly focusing on methods that allow LLMs to be adapted to new tasks while minimizing additional computational overhead and mitigating the associated carbon footprint.

  ▸ ***Mitigate:***

  Consequently, a diverse set of innovative solutions has been proposed to customize pre-trained language models for specialized tasks, ensuring seamless adaptation without extensive retraining. Low-Rank Adaptation (LoRA) (Hu et al., 2021) freezes the pre-trained model's weight matrices and injects trainable low-rank decomposition layers into each model layer. This approach is applicable to a wide range of models, effectively reduces memory requirements, and introduces no additional inference overhead—thereby enabling more efficient LLM adaptation. P-tuning (Liu et al., 2023) also keeps model weights fixed, appending a trainable embedding prefix to the input prompt to better tailor the model to the target task.

In addition, optimizing both training and alignment can reduce computational costs. Direct Preference Optimization (DPO) (Rafailov et al., 2024) eliminates the need for a separate reward model, improving performance while streamlining the training pipeline. The Odds Ratio Preference Optimization (ORPO) technique integrates supervised fine-tuning with preference alignment for further efficiency gains. Beyond algorithmic strategies, lower-level optimizations can also accelerate the training process. For instance, mixed-precision training (Micikevicius et al., 2018) diminishes memory usage and speeds up training, while the DeepSpeed framework (Rajbhandari et al., 2020) has been widely adopted for large-scale distributed LLM training to optimize memory and enhance throughput.

People have sophisticated metacognitive abilities in the area of memory ("metamemory"), both for themselves (knowing what they remember) and for other (knowing what others remember), which they can use in the context of transactive memory systems (e.g., asking a friend for a reminder about a shared experience). What might it look like to imbue AI agents with metamemory with respect to other agents (i.e., knowing what other agents remember)? In what contexts might it be useful? In what contexts might it be essential?
Answer:

Memory plays a vital role in our daily cognitive processes. It enables us to store learned experiences, associate them with emotions, and organize these experiences for efficient retrieval in the future. Given the critical importance of memory, recent research in large language model (LLM) agents has proposed integrating memory systems to enhance their ability to learn from past experiences and interact more effectively with their environment. For example, ReAct (Yao et al., 2023) allows LLM agents to reflect on past experiences to refine their action space, while Reflexion (Shinn et al., 2024) leverages episodic memory to guide the learning process. Metamemory, defined as an individual's knowledge and awareness of their own memory processes (Flavell & Wellman, 1975), has not yet been explicitly acknowledged or integrated into the current design of LLM agents.

Firstly, designing a transactive memory system, where agents can exchange what they know or understand, requires a new mechanism that enables agents to effectively articulate or summarize their knowledge. The current state-of-the-art agent systems implement memory structures based on the retrieval-augmented generation (RAG) framework (Park et al. (2023), Yu et al. (2024)). While RAG is effective (Fan et al., 2024) and offers flexibility in the choice of memory medium (Bag et al. (2024), Yang et al. (2024)), it is limited to answering "local" queries and cannot handle "global" queries, such as summarizing the general topic of the memory store. Without an additional mechanism, LLM agents face significant challenges in summarizing their memory stores—a critical first step in developing a functional transactive memory system. Edge et al. (2024) proposed a graph-based solution that summarizes the memory store by grouping memory events using a community detection algorithm. However, this approach is not a panacea for enabling agents to mimic human behavior, as the graph structure may fail to provide a viable method for summarizing memory in other forms, such as chronological events (Tenenbaum et al., 2011). Secondly, interchanging metamemory across all agents may be inefficient, as it demands significant computational power and introduces substantial redundancy. This inefficiency arises from the agents' inability to filter and attend to information with a focus. Thus, a "meta RAG" may be required as a central repository for metamemory, enabling individual agents to query summary information from others or directly connect with specific agents for collaboration and information exchange.

The transaction of metamemory holds the potential to create a system akin to a "hive mind," where knowledge is distributed and stored across multiple agents. In such a system, memory can be shared dynamically by establishing direct communication between agents, enabling seamless collaboration and information exchange when necessary. This approach would be particularly useful for building LLM agent systems that require strong collaboration. For example, in a coding LLM agent system for a large development project, individual agents could possess local knowledge about the code they produced while being coordinated to implement cross-module functionalities efficiently. Such a system may be essential in scenarios where aggregated responses are necessary. For instance, this approach could facilitate the formation of distinct groups of agents in financial market simulation use cases (Yao et al., 2024).

The word "bias" is used in many fields across the behavioral, social, and computational sciences. Learners have inductive biases. People have identity-based biases reflected in their implicit and explicit attitudes towards those of particular genders, races, ethnicities. People also have biases less directly connected to identity (e.g., a bias that grey-haired people are older). ML models can exhibit a bias–variance tradeoff. AI ethicists are concerned about bias inherent in NLP models, e.g., in embedding models. Neural networks can include a bias term. Define these various senses of "bias" (and any other senses that are commonly discussed in the literature on your reading list) and compare and contrast them. Do they all refer to the same concept? Do they all refer to distinct concepts?
Answer:

Bias, according to Merriam-Webster, refers to an inclination of temperament—often (though, as I will argue, misleadingly) linked to unfair judgment toward a particular person or thing. Beyond its social manifestations, the concept of "bias" appears in numerous scientific fields, each with its own nuances. Nevertheless, I propose that bias can be broadly understood as **an unconditionally systematic deviation from a neutral reference frame**. By itself, bias is neither inherently positive nor negative; rather, cultural context determines how we interpret it. In statistical learning, for instance, bias can improve modeling accuracy, though sometimes at the cost of increased variance. Conversely, in the realm of algorithmic fairness, social bias often necessitates mitigation to protect certain cultural subgroups. Despite its wide prevalence, bias can generally be categorized into the following types:

- **Bias in Statistical Machine Learning:**
  In statistics, bias refers to any feature of a statistical method that causes the estimated expected value of a population parameter to deviate systematically from its true value. Let the M denote the statistical method and P as the parameter(s) to be estimated, then the bias defined as

  $$a = \text{bias}(M, P) = E[P] - P$$

  where $E[\cdot]$ is the expectation function, and the a is the bias associated with the statistical M. While certain biases can be traced back to specific methods, others—such as sampling bias or omitted variable bias—are more general and can affect a wide range of statistical procedures.

  In the context of statistical learning, the concept of bias remains essentially unchanged; however, the focus shifts to estimating model parameters. Consequently, bias typically refers to the deviation of these estimated parameters from those of the true model. Additionally, the interplay between bias and variance is crucial to understanding how well a model generalizes to unseen data. This dynamic is captured by the well-known *bias–variance trade-off* (Pedro, 2000). Assuming a mean squared loss (MSE) objective, assuming the data is generated by the function $y = f(x) + \varepsilon$ where ε is random noise with mean 0 and variance δ, and denoting the model with estimated parameters as $\hat{f}(x)$, then the expected error on the unseen data can be decomposed into three parts

  $$E\Big[\Big(f(x) - \hat{f}(x)\Big)^2\Big] = \text{bias}\Big(\hat{f}(x)\Big)^2 + \text{var}\Big(\hat{f}(x)\Big)^2 + \delta$$

  Given a certain level of MSE, reducing variance inevitably increases bias, and vice versa. In particular, the bias term is closely tied to the "complexity" of the model; a simpler model (e.g., linear regression) may produce higher bias but exhibits more consistent performance across unseen examples (low variance). This principle underlies many regularization methods.

  Notably, the constant term in many machine learning algorithms is often referred to as the "bias" term (e.g., in neural networks or linear regression models). This notion of an unconditional deviation

persists because the constant remains independent of the inputs, shifting the model's outputs accordingly.

- **Social Bias:**
  The recent advances in auto-regressive transformer models have set new performance benchmarks across a range of tasks (Jiang et al. (2023), Qin et al. (2024), Qwen et al. (2025)). Departing from earlier approaches that favored specialized models for individual tasks, large language models (LLMs) now demonstrate strong capabilities in few-shot and zero-shot scenarios (Dong et al. (2024), Brown et al. (2020)). Furthermore, techniques like fine-tuning (Hu et al., 2021) and distillation (Groeneveld et al., 2024) enable these models to be efficiently deployed for numerous downstream applications and deployments. However, as these data-intensive models are typically pre-trained on large, web-crawled corpora, researchers have detected social biases that can discriminate against certain groups (Wan et al. (2023), Sheng et al. (2019)). Troublingly, findings also suggest that LLMs may even amplify such biases, indicating that the issue can become more pronounced over time (Benjamin, 2023).

  Following Gallegos et al. (2024), social bias is defined as "unfair discrimination encompassing disparate treatment or outcomes among social groups, arising from historical and structural power imbalances." In natural language processing (NLP), social bias can emerge in multiple tasks. Gallegos et al. (2024) offers a non-exhaustive list of the forms of the social bias in various tasks:
  - **Text Generation:** The social bias may be identified via the probability distribution of the next token.
  - **Machine Translation:** The translated word may default to masculine words.
  - **Question-Answering:** The answer may invoke stereotypes for a certain cultural group.
  - **Classification:** The classification results may biased toward certain cultural group.

  Given the variety of social biases documented by current research—for instance, gender bias (Bartl et al., 2020) and ethnic bias (Sap et al., 2019)—it is crucial to develop detection benchmarks and mitigation strategies to enhance the fairness of today's models. Barikeri et al. (2021) provides a conversational dataset grounded in real Reddit posts, enabling measurements of bias across gender, race, religion, and queerness. Additionally, Rudinger et al. (2018) offers a benchmark to evaluate gender bias by examining word associations for various social groups. To mitigate these biases and promote fairness, Qian et al. (2022) demonstrates that training on demographically perturbed corpora can lead to more equitable outcomes.

- **Cognitive Bias:**
  It has been shown that people routinely use heuristics in their daily lives, which sheds light on many of their decisions and behaviors (Griffiths & Tenenbaum, 2006). Interestingly, neural networks trained via optimization algorithms have likewise been observed to rely on heuristics—often to good effect for in-distribution data, though not necessarily in a universally optimal manner (McCoy et al., 2019). Consequently, while these heuristics are not guaranteed to be correct or fully generalizable, they underpin much of our day-to-day cognitive processes. Although some define cognitive biases as "systematic patterns of deviation from rational judgment," I contend that the biases should be understood as heuristics, which are not necessarily bound to rationality. Such heuristics may develop from accumulated experience and, owing to their shortcut-like nature, can lead to irrational behaviors or understandings. Moreover, as social interactions play a crucial role in shaping our experiences, these heuristics may also manifest as social biases. Given that neural networks learn in a comparable way —adapting to patterns in available data—it is unsurprising that both humans and neural networks exhibit similar biases (Caliskan et al., 2017).

In this sense, cognitive biases—or heuristics—function as a prior for understanding, shaping the way we interpret and integrate new information. Building on experimental and modeling work, Oosterhof & Todorov (2008) shows that impressions of faces can be decomposed into two orthogonal dimensions: valence and dominance. Valence indicates whether we are inclined to approach or avoid someone, whereas dominance relates to the person's perceived physical strength. Consequently, affective traits such as sadness, anger, or happiness can be represented as combinations of these two features. Furthermore, Peterson et al. (2022) demonstrates that high-dimensional latent feature vectors for human faces, derived via StyleGAN (Karras et al., 2019), can be leveraged to create a model closely aligning with human perception. Manipulating these latent vectors yields changes in facial photographs that remain consistent with human evaluations. Meanwhile, certain cognitive biases are intrinsically tied to notions of rational decision-making, further illustrating how shortcuts in judgment can affect our capacity for logical analysis (Tversky & Kahneman, 1974). Interestingly, Jones & Steinhardt (2022) finds that these irrational biases are not limited to humans, as they also emerge in LLMs.

For the rest of the question, in my opinion, the biases all these biases involve a systematically unconditional deviation from a reference frame to some degree, yet each operates under its own distinct context.

Bibliography

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., … Zhou, X. (2024,). *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*. https://arxiv.org/abs/2404.14219

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, And Society*, 298–306.

Acharya, A., Singh, B., & Onoe, N. (2023). Llm based generation of item-description for recommendation system. *Proceedings of the 17th ACM Conference on Recommender Systems*, 1204–1207.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., & others. (2023). Gpt-4 technical report. *Arxiv Preprint Arxiv:2303.08774*.

Alain, G., & Bengio, Y. (2018). Understanding Intermediate Layers Using Linear Classifier Probes. 2018. *Arxiv Preprint Arxiv:1610.01644*.

Anonymous. (2024,). Identifying and Tuning Safety Neurons in Large Language Models. *Submitted to the Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=yR47RmND1m

Bag, S., Gupta, A., Kaushik, R., & Jain, C. (2024). RAG Beyond Text: Enhancing Image Retrieval in RAG Systems. *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET*, *0*, 1–6. https://doi.org/10.1109/ICECET61485.2024.10698598

Barikeri, S., Lauscher, A., Vulić, I., & Glavaš, G. (2021). RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers): Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2021.acl-long.151

Bartl, M., Nissim, M., & Gatt, A. (2020). Unmasking Contextual Stereotypes: Measuring and Mitigating BERT`s Gender Bias. In M. R. Costa-jussà, C. Hardmeier, W. Radford, & K. Webster (Eds.), *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing: Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. https://aclanthology.org/2020.gebnlp-1.1/

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Benjamin, R. (2023). Race after technology. In *Social Theory Re-Wired: Social Theory Re-Wired* (pp. 405–415). Routledge.

Bereska, L., & Gavves, E. (2024). Mechanistic Interpretability for AI Safety–A Review. *Arxiv Preprint Arxiv:2404.14082*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. von, Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022,). *On the Opportunities and Risks of Foundation Models*. https://arxiv.org/abs/2108.07258

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., … Olah, C. (2023). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020,). *Language Models are Few-Shot Learners*. https://arxiv.org/abs/2005.14165

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Chen, H., Zhu, T., Liu, B., Zhou, W., & Philip, S. Y. (2024). Fine-tuning a Biased Model for Improving Fairness. *IEEE Transactions on Big Data*.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., & others. (2021). Training verifiers to solve math word problems. *Arxiv Preprint Arxiv:2110.14168*.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. *Arxiv Preprint Arxiv:2309.08600*.

Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., & Wei, F. (2021). Knowledge neurons in pretrained transformers. *Arxiv Preprint Arxiv:2104.08696*.

De Sabbata, C. N., Sumers, T. R., & Griffiths, T. L. (2024). Rational metareasoning for large language models. *Arxiv Preprint Arxiv:2410.05563*.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., … Pan, Z. (2024,). *DeepSeek-V3 Technical Report*. https://arxiv.org/abs/2412.19437

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers): Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. https://doi.org/10.18653/v1/N19-1423

Ding, Y., & Shi, T. (2024). Sustainable LLM Serving: Environmental Implications, Challenges, and Opportunities : Invited Paper. *2024 IEEE 15th International Green and Sustainable Computing Conference (IGSC)*, *0*, 37–38. https://doi.org/10.1109/IGSC64514.2024.00016

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Liu, T., Chang, B., Sun, X., Li, L., & Sui, Z. (2024,). *A Survey on In-context Learning*. https://arxiv.org/abs/2301.00234

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021,). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. https://arxiv.org/abs/2010.11929

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2024,). *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*. https://arxiv.org/abs/2404.16130

Elad, M. (2010). *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media.

Elhage, N., Hume, T., Olsson, C., Nanda, N., Henighan, T., Johnston, S., ElShowk, S., Joseph, N., DasSarma, N., Mann, B., Hernandez, D., Askell, A., Ndousse, K., Jones, A., Drain, D., Chen, A., Bai, Y., Ganguli, D., Lovitt, L., … Olah, C. (2022b). Softmax Linear Units. *Transformer Circuits Thread*.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., & others. (2022a). Toy models of superposition. *Arxiv Preprint Arxiv:2209.10652*.

Engels, J., Michaud, E. J., Liao, I., Gurnee, W., & Tegmark, M. (2024,). *Not All Language Model Features Are Linear*. https://arxiv.org/abs/2405.14860

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024). A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6491–6501. https://doi.org/10.1145/3637528.3671470

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning*, 1126–1135.

Flavell, J. H., & Wellman, H. M. (1975). *Metamemory*.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–79.

Gentner, D., & Stevens, A. L. (2014). *Mental models*. Psychology Press.

Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., & Berant, J. (2021). Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, *9*, 346–361.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, *29*, 24–30.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K. R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., … Hajishirzi, H. (2024,). *OLMo: Accelerating the Science of Language Models*. https://arxiv.org/abs/2402.00838

Hao, S., Gu, Y., Luo, H., Liu, T., Shao, X., Wang, X., Xie, S., Ma, H., Samavedhi, A., Gao, Q., & others. (2024). LLM Reasoners: New Evaluation, Library, and Analysis of Step-by-Step Reasoning with Large Language Models. *Arxiv Preprint Arxiv:2404.05221*.

Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., & Hu, Z. (2023). Reasoning with language model is planning with world model. *Arxiv Preprint Arxiv:2305.14992*.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4129–4138.

Hong, J., Lee, N., & Thorne, J. (2024,). *ORPO: Monolithic Preference Optimization without Reference Model*. https://arxiv.org/abs/2403.07691

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021,). *LoRA: Low-Rank Adaptation of Large Language Models*. https://arxiv.org/abs/2106.09685

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., & Khabsa, M. (2023,). *Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations*. https://arxiv.org/abs/2312.06674

Ioffe, S., & Szegedy, C. (2015,). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. https://arxiv.org/abs/1502.03167

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, *120*(11). https://doi.org/10.1073/pnas.2208839120

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

Jin, X., Barbieri, F., Kennedy, B., Davani, A. M., Neves, L., & Ren, X. (2020). On transferability of bias mitigation effects in language model fine-tuning. *Arxiv Preprint Arxiv:2010.12864*.

Jones, E., & Steinhardt, J. (2022). Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, *35*, 11785–11799.

Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Arxiv Preprint Arxiv:1705.03551*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *Arxiv Preprint Arxiv:2001.08361*.

Karras, T., Laine, S., & Aila, T. (2019,). *A Style-Based Generator Architecture for Generative Adversarial Networks*. https://arxiv.org/abs/1812.04948

Kim, J., Park, S., Kwon, Y., Jo, Y., Thorne, J., & Choi, E. (2023,). *FactKG: Fact Verification via Reasoning on Knowledge Graphs*. https://arxiv.org/abs/2305.06590

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016,). *Building Machines That Learn and Think Like People*. https://arxiv.org/abs/1604.00289

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Li, T., Khashabi, D., Khot, T., Sabharwal, A., & Srikumar, V. (2020). UNQOVERing Stereotyping Biases via Underspecified Questions. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020: Findings of the Association for Computational Linguistics: EMNLP 2020*. https://doi.org/10.18653/v1/2020.findings-emnlp.311

Liu, A., Pan, L., Lu, Y., Li, J., Hu, X., Zhang, X., Wen, L., King, I., Xiong, H., & Yu, P. S. (2024,). *A Survey of Text Watermarking in the Era of Large Language Models*. https://arxiv.org/abs/2312.07913

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & Tang, J. (2023,). *GPT Understands, Too*. https://arxiv.org/abs/2103.10385

Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2022,). *Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model*. https://arxiv.org/abs/2211.02001

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019,). *On Measuring Social Biases in Sentence Encoders*. https://arxiv.org/abs/1903.10561

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., & Hendrycks, D. (2024,). *HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal*. https://arxiv.org/abs/2402.04249

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/P19-1334

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., & Wu, H. (2018,). *Mixed Precision Training*. https://arxiv.org/abs/1710.03740

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023,). *DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature*. https://arxiv.org/abs/2301.11305

Mitrović, S., Andreoletti, D., & Ayoub, O. (2023,). *ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text*. https://arxiv.org/abs/2301.13852

Mondal, D., & Lipizzi, C. (2024). Mitigating Large Language Model Bias: Automated Dataset Augmentation and Prejudice Quantification. *Computers*, *13*(6), 141.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022,). *Training language models to follow instructions with human feedback*. https://arxiv.org/abs/2203.02155

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual Acm Symposium on User Interface Software and Technology*, 1–22.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. (2022). BBQ: A hand-built bias benchmark for question answering. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Findings of the Association for Computational Linguistics: ACL 2022: Findings of the Association for Computational Linguistics: ACL 2022*. https://doi.org/10.18653/v1/2022.findings-acl.165

Pedro, D. (2000). A unified bias-variance decomposition and its applications. *17th International Conference on Machine Learning*, 231–238.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018,). *Deep contextualized word representations*. https://arxiv.org/abs/1802.05365

Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, *119*(17), e2115228119.

Qian, R., Ross, C., Fernandes, J., Smith, E. M., Kiela, D., & Williams, A. (2022). Perturbation Augmentation for Fairer NLP. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2022.emnlp-main.646

Qin, Y., Li, X., Zou, H., Liu, Y., Xia, S., Huang, Z., Ye, Y., Yuan, W., Liu, H., Li, Y., & Liu, P. (2024,). *O1 Replication Journey: A Strategic Progress Report – Part 1*. https://arxiv.org/abs/2410.18982

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., … Qiu, Z. (2025,). *Qwen2.5 Technical Report*. https://arxiv.org/abs/2412.15115

Radford, A., & Narasimhan, K. (2018,). *Improving Language Understanding by Generative Pre-Training*. https://api.semanticscholar.org/CorpusID:49313245

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2024,). *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. https://arxiv.org/abs/2305.18290

Rajbhandari, S., Rasley, J., Ruwase, O., & He, Y. (2020,). *ZeRO: Memory Optimizations Toward Training Trillion Parameter Models*. https://arxiv.org/abs/1910.02054

Rajpurkar, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *Arxiv Preprint Arxiv:1606.05250*.

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). Gender Bias in Coreference Resolution. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers): Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. https://doi.org/10.18653/v1/N18-2002

Santurkar, S., Tsipras, D., Ilyas, A., & Madry, A. (2018). How does batch normalization help optimization?. *Advances in Neural Information Processing Systems*, *31*.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678.

Saparov, A., & He, H. (2022). Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *Arxiv Preprint Arxiv:2210.01240*.

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019,). *The Woman Worked as a Babysitter: On Biases in Language Generation*. https://arxiv.org/abs/1909.01326

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2024). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, *36*.

Shridhar, M., Yuan, X., Côté, M.-A., Bisk, Y., Trischler, A., & Hausknecht, M. (2020). Alfworld: Aligning text and embodied environments for interactive learning. *Arxiv Preprint Arxiv:2010.03768*.

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances*, *9*(26), eadh1850. https://doi.org/10.1126/sciadv.adh1850

Sturmfels, P., Lundberg, S., & Lee, S.-I. (2020). Visualizing the Impact of Feature Attribution Baselines. *Distill*. https://doi.org/10.23915/distill.00022

Sundararajan, M., Taly, A., & Yan, Q. (2017,). *Axiomatic Attribution for Deep Networks*. https://arxiv.org/abs/1703.01365

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., & Metzler, D. (2020,). *Long Range Arena: A Benchmark for Efficient Transformers*. https://arxiv.org/abs/2011.04006

Templeton, A. (2024). *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Timor, N., Mamou, J., Korat, D., Berchansky, M., Pereg, O., Wasserblat, M., Galanti, T., Gordon, M., & Harel, D. (2024,). *Distributed Speculative Inference of Large Language Models is Provably Faster*. https://arxiv.org/abs/2405.14105

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2021,). *Universal Adversarial Triggers for Attacking and Analyzing NLP*. https://arxiv.org/abs/1908.07125

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., & Peng, N. (2023). " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *Arxiv Preprint Arxiv:2310.09219*.

Wang, A. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *Arxiv Preprint Arxiv:1804.07461*.

Wang, X., & Zhou, D. (2024). Chain-of-thought reasoning without prompting. *Arxiv Preprint Arxiv:2402.10200*.

Wang, Y., Zhang, Z., Zhang, P., Yang, B., & Wang, R. (2024,). *Meta-Reasoning: Semantics-Symbol Deconstruction for Large Language Models*. https://arxiv.org/abs/2306.17820

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., & others. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021,). *Ethical and social risks of harm from Language Models*. https://arxiv.org/abs/2112.04359

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

Winograd, T. (1971). *Procedures as a representation for data in a computer program for understanding natural language*.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023,). *BloombergGPT: A Large Language Model for Finance*. https://arxiv.org/abs/2303.17564

Xia, S., Li, X., Liu, Y., Wu, T., & Liu, P. (2024,). *Evaluating Mathematical Reasoning Beyond Accuracy*. https://arxiv.org/abs/2404.05692

Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K., & Wang, D. (2024). Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *8*(1), 1–32.

Yang, K., & Menczer, F. (2024). Anatomy of an AI-powered malicious social botnet. *Journal of Quantitative Description: Digital Media*, *4*. https://doi.org/10.51685/jqd.2024.icwsm.7

Yang, M., Shi, B., Le, M., Hsu, W.-N., & Tjandra, A. (2024,). *Audiobox TTA-RAG: Improving Zero-Shot and Few-Shot Text-To-Audio with Retrieval-Augmented Generation*. https://arxiv.org/abs/2411.05141

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, *36*.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *Arxiv Preprint Arxiv:2210.03629*.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023,). *ReAct: Synergizing Reasoning and Acting in Language Models*. https://arxiv.org/abs/2210.03629

Yao, Z., Li, Z., Thomas, M., & Florescu, I. (2024,). *Reinforcement Learning in Agent-Based Market Simulation: Unveiling Realistic Stylized Facts and Behavior*. https://arxiv.org/abs/2403.19781

Yu, Y., Li, H., Chen, Z., Jiang, Y., Li, Y., Zhang, D., Liu, R., Suchow, J. W., & Khashanah, K. (2024). FinMem: A performance-enhanced LLM trading agent with layered memory and character design. *Proceedings of the AAAI Symposium Series*, *3*(1), 595–597.

Zeng, Y., Yang, Y., Zhou, A., Tan, J. Z., Tu, Y., Mai, Y., Klyman, K., Pan, M., Jia, R., Song, D., Liang, P., & Li, B. (2024,). *AIR-Bench 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies*. https://arxiv.org/abs/2407.17436

Zeng, Z., Chen, P., Liu, S., Jiang, H., & Jia, J. (2024b,). *MR-GSM8K: A Meta-Reasoning Benchmark for Large Language Model Evaluation*. https://arxiv.org/abs/2312.17080

Zeng, Z., Liu, Y., Wan, Y., Li, J., Chen, P., Dai, J., Yao, Y., Xu, R., Qi, Z., Zhao, W., & others. (2024a). MR-BEN: A Comprehensive Meta-Reasoning Benchmark for Large Language Models. *Arxiv Preprint Arxiv:2406.13975*.

Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., & Li, L. (2023). Multilingual machine translation with large language models: Empirical results and analysis. *Arxiv Preprint Arxiv:2304.04675*.

Zhu, X., Li, J., Liu, Y., Ma, C., & Wang, W. (2024). A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, *12*, 1556–1577.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023,). *Universal and Transferable Adversarial Attacks on Aligned Language Models*. https://arxiv.org/abs/2307.15043