

Position: On the Methodological Pitfalls of Evaluating Base LLMs for Reasoning

Jason Chan

Zhixue Zhao

Robert Gaizauskas

University of Sheffield, UK

{jlychan1, r.gaizauskas, zhixue.zhao}@sheffield.ac.uk

Abstract

Existing work investigates the reasoning capabilities of large language models (LLMs) to uncover their limitations, human-like biases and underlying processes. Such studies include evaluations of base LLMs (pre-trained on unlabeled corpora only) for this purpose. Our position paper argues that evaluating base LLMs’ reasoning capabilities raises inherent methodological concerns that are overlooked in such existing studies. We highlight the fundamental mismatch between base LLMs’ pretraining objective and normative qualities, such as correctness, by which reasoning is assessed. In particular, we show how base LLMs generate logically valid or invalid conclusions as coincidental byproducts of conforming to purely linguistic patterns of statistical plausibility. This fundamental mismatch challenges the assumptions that (a) base LLMs’ outputs can be assessed as their *bona fide* attempts at correct answers or conclusions; and (b) conclusions about base LLMs’ reasoning can generalize to post-trained LLMs optimized for successful instruction-following. We call for a critical re-examination of existing work that relies implicitly on these assumptions, and for future work to account for these methodological pitfalls.

1 Introduction

Reasoning is the process of drawing conclusions based on existing information (Holyoak and Morrison, 2012; Mondorf and Plank, 2024). It is a cornerstone of intelligence that has traditionally been the subject of extensive study in philosophy and psychology. In artificial intelligence, systems as early as the Logic Theorist (Newell and Simon, 1956) and General Problem Solver (Newell et al., 1959) have been proposed to simulate human reasoning processes using rules in formal logic.

This longstanding pursuit of artificial systems that can simulate reasoning continues to this day, as the reasoning capabilities and behavior of large language models (LLMs) have grown into a highly active area of current research (Mondorf and Plank, 2024; Ke et al., 2025, etc.). However, in contrast to earlier symbolic systems designed expressly to simulate reasoning, LLMs’ ability to reason has been characterized as an “*emergent*” property (Wei et al., 2022b; Berti et al., 2025)¹ that can be “*elicited*” by prompting (Wei et al., 2022c).

Accordingly, current work commonly rests on a key assumption: *the reasoning behavior and performance of an LLM can be evaluated by assessing its outputs in response to the appropriate prompts, regardless of whether the model itself has undergone any form of post-training to optimize it for instruction-following and correct reasoning*. On this basis, comparisons among different model sizes and types (e.g. pre-trained only versus instruction-tuned) are typically treated as only a matter of degree in capability, e.g. with instruction-tuning often described as a means of *improving* a model’s reasoning capabilities (Chung et al., 2024). This premise underlies a wide range of studies that have experimented wholly or partly with base LLMs to draw conclusions about systemic reasoning errors, human-like biases, and specific “circuits” believed to explain a model’s reasoning behavior (see Section 3).

In this position paper, we challenge this foundational assumption and argue that **the practice of evaluating base LLMs for reasoning raises a number of inherent methodological issues overlooked in existing work**. Specifically:

1. We highlight a fundamental mismatch in that base LLMs are optimized for linguistic plausibility instead of correctness or other similar

¹But see Schaeffer et al. (2023) and Lu et al. (2024) etc. for differing views.

- desirable normative qualities that reasoning tasks are designed to evaluate (Section 4);
2. We argue and empirically demonstrate how base LLMs conforming to simple linguistic patterns can generate logically valid or invalid conclusions as purely coincidental byproducts of optimizing for linguistic plausibility (Section 5);
 3. We argue that the ambiguity of whether base LLMs’ outputs are *bona fide* attempts at correct answers is a critical confounding variable in assessing these models for reasoning abilities. When uncontrolled, this confound undermines conclusions about the models’ inherent reasoning limitations or systematic biases (Section 6); and
 4. Finally, we argue that findings on base LLMs’ reasoning cannot be assumed to generalize to instruct LLMs, as they optimize for fundamentally different objectives. We therefore call for such claims in existing studies to be critically re-examined (Section 6.3).

2 Training LLMs

A typical LLM training pipeline (e.g. Grattafiori et al., 2024; Abdin et al., 2024; Lambert et al., 2025) broadly consists of two stages: pre-training and post-training.

In **pre-training**, a model is trained on large text corpora to maximize the conditional probability of the next token given its preceding context. These corpora include general web pages, textbooks, academic papers, code repositories and synthetically generated datasets, with web pages typically occupying a large portion of the data mix (e.g. Biderman et al., 2023; OLMo et al., 2025).

In certain cases, pre-training itself is split into two sub-stages, such that, in the latter sub-stage (commonly referred to as “*mid-training*”), the model is trained on a smaller subset of unlabeled data that is typically domain-specific (e.g. math, code) and of a higher quality (see e.g. Feng et al., 2024; Blakeney et al., 2024; OLMo et al., 2025).

We refer to models that have been pre-trained only, without any post-training, as *base LLMs*.

Broadly speaking, the goal of **post-training** is to optimize a base LLM to follow user instructions, which includes correctly answering questions and successfully completing tasks specified

in prompts to the model. For this purpose, post-training typically consists of multiple sub-stages, such as instruction-tuning (IT) (Wei et al., 2022a) and preference-tuning (PT) (Ouyang et al., 2022), which differ among model families.

In IT, the model is likewise trained to maximize the conditional probability of the next token given its preceding context, but with three key differences illustrated by the following sequence²:

```
<|endoftext|><|user|>\n Explain
University level Introductory
Statistics to me like I'm a
child\n <|assistant|>\nOkay
little buddy, imagine you have a
big bag of differently colored
candies ... <|endoftext|>
```

1. Each instance in the IT data is a pair of text sequences: a **user instruction** and a **target response** to that instruction.³
2. Each pair of text sequences is formatted with a chat template (which demarcates instructions with special tokens e.g. “<|user|>”) and concatenated into a single sequence.
3. Only the loss corresponding to tokens in the **target response** sequence is taken into account, whereas loss corresponding to the **user instruction** sequence is masked.

In PT, the model is further optimized using datasets with each instance consisting of an instruction prompt, a preferred response and a rejected response (see e.g. Bai et al., 2022). While various techniques such as direct preference optimization (Rafailov et al., 2023) are introduced for this purpose, a common objective is for the model to maximize the difference between the log likelihood assigned to the preferred response and that assigned to the rejected response. This step ensures that outputs generated by the model are desirable according to human judgments⁴ based on a specified set of normative criteria, such as “*helpfulness*”, “*instruction following*”, “*honesty*” and “*truthfulness*” (Lambert et al., 2025).

We refer to models that have undergone a post-training stage⁵ as *instruct LLMs*.

²Example from Zhao et al. (2024).

³For our purposes, we can disregard the use of system prompts and multi-turn sequences.

⁴Or, in LLM-as-a-judge (Zheng et al., 2023) or RLAIF (Lee et al., 2024a) approaches, judgments of existing LLMs that have themselves already been preference-tuned.

⁵We interpret *post-training* broadly to include any method

3 Current Work in LLM Reasoning

The practice of evaluating base LLMs on reasoning is prevalent in existing work. As we show in this Section with representative studies, this is typically carried out with two implicit assumptions:

1. When prompted, base LLMs are genuinely attempting to answer correctly as per the task instructions; and
2. Findings from base LLMs (e.g. on their reasoning limitations) can be generalized to support claims about LLMs as a whole, including instruct LLMs.

While we will challenge both these assumptions in subsequent Sections, we first establish the context for our critique by illustrating its necessity in light of current research practices.

Capability evaluation. Numerous studies have evaluated LLMs’ ability to reason with logic (Liu et al., 2025), mathematics (Ahn et al., 2024), commonsense (Li et al., 2022; Chan et al., 2025), specialist and scientific knowledge (Zhang et al., 2024) across various domains (see e.g. Yu et al., 2024 and Giadikiaroglou et al., 2024 for further surveys). As shown in Table 1, studies typically assume that reasoning can be elicited from base LLMs as well as instruct LLMs, and that their capabilities differ only in degree (i.e. which type of LLM reasons *better*) and not in nature (see e.g. Li et al., 2025). Based on this assumption, studies have also conducted experiments exclusively on base LLMs to draw conclusions about reasoning limitations of LLMs *in general* (see Table 1).

Comparisons with human reasoning. Similarly, the assumption that base LLMs are reasoning and genuinely attempting to generate correct answers or conclusions is also relied upon by studies that compare the reasoning behavior of LLMs against that of human reasoners. Existing work in this category has found that LLMs exhibit certain human-like cognitive biases, such as content effect (Lampinen et al., 2024), base rate fallacy (Yax et al., 2024) and figural effect (Eisape et al., 2024).

Behavioral understanding. A growing body of interpretability research aims to understand the

of training whose goal is to optimize a model to follow user instructions in some desirable manner e.g. successfully, helpfully etc., which includes techniques such as applying reinforcement learning (with correct answers as reward signals) directly to base LLMs (DeepSeek-AI et al., 2025) without going through any fine-tuning stages such as IT as we described.

internal processes behind LLM reasoning. One commonly adopted approach is mechanistic interpretability (Olah et al., 2020), which aims to identify specific computational subgraphs (or “circuits”) that causally explain how a model computes its outputs when prompted with certain contexts or premises (e.g. Wang et al., 2023; Hong et al., 2024; Nikankin et al., 2025). As shown in Table 1, such mechanistic analyses have typically been conducted only on base LLMs.

Circuits are typically identified using *activation patching* (Vig et al., 2020; Zhang and Nanda, 2024), which involves pairs of minimally differing prompts, *clean* versus *corrupt*. For example, to identify the circuit responsible for a model’s capability for syllogistic reasoning, Kim et al. (2025) uses clean prompts in the form of “*All A are B. All B are C. Therefore, all A are*”⁶ and corrupt prompts “*All A are B. All Q are C. Therefore, all A are*”. First, the model is run on both the clean and corrupt prompts, with its respective output logits recorded. In the clean run, intermediate output values of selected model components (e.g. individual attention heads at different model layers) are also stored. Then, when the model is run again on the corrupt prompt, the value of a particular component is overwritten by “patching in” the corresponding stored value from the clean run, and the resulting change in the model’s output logits is recorded. By systematically patching each model component, researchers can identify those that result in large changes, which would indicate that these components play a key role in the circuit for the particular model behavior being studied.

4 The Fundamental Mismatch in Evaluating Base LLMs on Reasoning

Our position recognizes but sidesteps the contentious theoretical debate on whether LLMs are capable of *genuine* reasoning, and the philosophical issues with characterizing what this entails (see e.g. Shanahan, 2024; Cappelen and Dever, 2025). Instead, we focus on the practical issue of *evaluation methodology*. Specifically, we argue that the practice of experimenting on base LLMs by prompting in order to draw conclusions about the reasoning limitations and processes of LLMs *in general* (whether base or instruct) is problematic.

⁶In this case, a model capable of syllogistic reasoning is expected to predict “C” as the most likely next token.

Category	Subcategory	Study	Reasoning domain/type	Base LLMs?	Instruct LLMs?
Evaluating capabilities and limitations	Comparative analysis (base vs instruct)	Li et al., 2025	General	✓	✓
		Bigoulaeva et al., 2025	General	✓	✓
	Comparative analysis (model scale)	Chen et al., 2024	General; math	✓	✗
		McKenzie et al., 2023	General	✓	✓
	Benchmark evaluation	Srivastava et al., 2024	Math	✓	✓
		Jin et al., 2023	Causal	✓	✓
		Yang et al., 2024	Inductive	✓	✗
	Benchmark evaluation; error analysis	Griot et al., 2025	Medical	✓	✓
	Error analysis	(Ye and Durrett, 2022)	General	✓	✓
		Zheng et al., 2024	General	✓	✓
		Razeghi et al., 2022	Math	✓	✗
Comparisons with human reasoning	Enhancement method	Huang et al., 2023	General	✓	✗
		Sharma et al., 2024	General	✓	✗
		Hao et al., 2025	General	✓	✗
	Cognitive bias and error analysis	Lampinen et al., 2024	General	✓	✓
		Yax et al., 2024	General	✓	✓
		Binz and Schulz, 2023	General	✓	✓
		Eisape et al., 2024	Logic	✓	✗
	Mechanistic interpretability	Hou et al., 2023	General; CoT	✓	✗
		Dutta et al., 2024	Logic; CoT	✓	✗
		Kim et al., 2025	Logic	✓	✗
		Stolfo et al., 2023	Math	✓	✗
		Lee et al., 2024b	Causal	✓	✗

Table 1: Examples of existing representative work in LLM reasoning. As shown in the second rightmost column, all the listed studies conduct experiments with base LLMs on the **assumption that these models can be prompted to engage in reasoning**. As shown in the rightmost column, relying on this assumption, some studies compare the reasoning of base LLMs with instruct LLMs, while others conduct experiments exclusively using base LLMs.

4.1 Core Principle: Performance Measures should Reflect Models’ Objectives

A core principle in machine learning is that the performance measures on which a model is evaluated should reflect the tasks that the model is learning to solve (Mitchell, 1997; Goodfellow et al., 2016). For example, consider a set of medical image classifiers designed and trained only for classifying benign versus malignant tumors. If we evaluate these models on how accurately they predict the gender of the patients, we are evaluating them on a task they were never optimized for. Suppose we find these models perform far worse at predicting the gender of older patients than younger ones. Can we conclude that this type of classifier generally exhibits bias in gender prediction?

We argue that this conclusion is unwarranted. A prediction bias is a systematic error **as measured against the model’s targets** and not against an arbitrary variable that the model was not trained to predict in the first place. In this case, there is a fundamental mismatch between the task being evaluated (gender prediction) and the model’s objective (tumor classification). In other words, any capa-

bility the model has to predict gender is *purely incidental to its objective*. As such, the finding cannot support the conclusion that, even if models of this type were *specifically* trained or fine-tuned for gender prediction, they would exhibit the same systematic bias in performing worse at predicting the gender of older patients than younger ones.

4.2 Base LLMs Optimize for Linguistic Plausibility, not for Normative Qualities

As described in Section 2, a base LLM’s objective optimizes only for linguistic plausibility and not for normative qualities, such as correctness or logical validity, by which we typically assess reasoning. In other words, base LLMs are not trained for the purpose of producing outputs that conform to our standards for what counts as “good” reasoning; instead, they are trained only for the purpose of producing outputs that are statistically the most likely continuation of a given context.⁷

Applying the analogy in Section 4.1, the heart of our argument is that **the ability of base LLMs to produce outputs that reflect good reasoning**

⁷We use the term “linguistic plausibility” in this specific purely statistical sense (as opposed to e.g. grammaticality).

is therefore purely incidental to its objective in optimizing for linguistic plausibility.

4.3 Why High-Quality Pretraining Data does not Address the Mismatch

Here, a key objection to this position is that by curating high-quality pretraining corpora, base LLMs are already *implicitly* optimizing for correctness and other desirable normative qualities associated with reasoning. We argue that this view misinterprets both the nature of high-quality data and the objective of the base LLMs.

While corpora typically considered as high-quality (such as human-written academic textbooks) can consist of example questions (or premises) followed by correct answers (or valid conclusions), they can also contain *bona fide* demonstrations of incorrect answers and flawed reasoning. This is by no means an issue of poor data quality, misinformation or bias, the effects of which have all been discussed in existing work (e.g. Lin et al., 2022; McKenzie et al., 2023). On the contrary, these demonstrations of incorrect answers and reasoning are a valuable feature in high-quality educational texts that serve a clear pedagogical purpose.⁸

For instance, a textbook might include demonstrations in the format of “*Question: [...] Answer: [...]*”, followed by a helpful explanation of why the answer is incorrect. Reproducing this purely statistical pattern, a base LLM prompted with a multiple-choice question might continue the context “*Question: [...] Answer: ”* not with what it “believes” to be correct answer, but with an *incorrect* answer accompanied by an explanation, e.g. “*C. The answer is incorrect because...”*.

What this shows is that pre-training a model on high-quality data does not on its own warrant the claim that the model’s outputs are therefore optimizing for correctness or validity as opposed to mere linguistic plausibility.

⁸In the pretraining corpus olmo-mix-1124 (OLMo et al., 2025), for example, we identified extracts of **incorrect** reasoning processes intended as demonstrations of e.g. a “*common mistake in applying the demand and supply framework*” (proofpile-shard-0035-31144), and how an incorrect solution to a “*Bayesian inference problem*” might be mistakenly derived (proofpile-shard-0035-24997).

4.4 Correctness and Normative Qualities are Valid Measures for Instruct LLMs

As recognized by Ouyang et al. (2022), the task of language modeling (predicting the most likely continuation of a preceding context, given an unlabeled pretraining corpus) is distinct from the task of helpfully and safely following user instructions. This distinction is what motivates instruction-tuning, so that models like Instruct-GPT (Ouyang et al., 2022) are re-optimized for the latter task. We argue that this re-optimization is precisely what entitles us to evaluate instruct LLMs for normative qualities such as correctness in their generated outputs.

For example, as explained in Section 2, instruct LLMs are fine-tuned on instruction-response pairs where the response is expected to *always* demonstrate a *successful fulfillment* of the instruction. This data is formatted with a chat template using unique markers such as “<|user|>” and “<|assistant|>” that are not expected to have appeared anywhere in the model’s pretraining data.

Because the text following “<|assistant|>” is always a demonstration of *successful* instruction-following, we are deliberately signaling to the model that its own generated continuation must also be a successful fulfillment. This includes correctly answering a question or drawing a logical conclusion, as opposed to simply demonstrating an unsuccessful response.⁹

This functional shift from language modeling to instruction-following means that generating correct and logical answers is no longer an incidental byproduct but a valid and important performance measure for how well a model optimizes for its new objective. Just as humans apply reasoning to achieve goals in real-life scenarios (Evans et al., 1993), reasoning plays a clear instrumental role in enabling an instruct LLM to successfully fulfill a wide variety of user instructions, such as correctly answering complex questions or debugging code (Ahmad et al., 2024). While the model is still mechanically performing next-token prediction, its re-optimization is intended to ensure that what the model considers to be linguistically plausible now **encapsulates** the normative qualities, such as correctness and help-

⁹As discussed in Section 2, we recognize that such fine-tuning is only one of the ways to re-optimize the model for correctness and provide it with a clear signal to generate a correct answer.

fulness, that we expect from responses that successfully fulfill user instructions.

5 Case Study: Debunking the Emergence of Logical Reasoning in Base LLMs

We present a case study to explain how logically valid or invalid conclusions can emerge as coincidental byproducts from a base LLM optimized only for linguistic plausibility. We show how observations about these byproducts can lead to unwarranted claims about the model’s underlying reasoning capabilities and limitations, demonstrating the methodological issues we set out in the previous Section.

5.1 Mechanisms of Linguistic Plausibility

Existing work has identified a number of behavioral patterns in base LLMs and the circuits that mechanically explain such behaviors (Olsson et al., 2022; Wang et al., 2023; Hanna et al., 2023 etc.). We isolate two circuits in particular:

The **Indirect Object Identification** (IOI) circuit (Wang et al., 2023) explains how a model predicts that a context in the form of “*When John and Mary went to the store, John gave a bottle to*” should be followed by “*Mary*”. The circuit implementing this behavior:

1. identifies all the names that have appeared in the context (“*Mary*” and “*John*”);
2. suppresses the name which has already been repeated in the context (“*John*”); and
3. outputs the remaining name (“*Mary*”).

An **induction head** (IH) circuit (Olsson et al., 2022) explains how a model predicts that a context of the form “[A]/[B]....[A]” should be followed by “[B]”. For example, if two tokens “*Harry Potter*” have appeared in the context, the model predicts that the next time the token “*Harry*” appears, it will likely be followed again by the token “*Potter*”.

As Figure 1 shows, the circuit for this behavior:

1. identifies the previous occurrence of a token “[A]” in the context (*prefix-matching*); and
2. boosts the logit of the token that followed “[A]” when “[A]” previously occurred (i.e. “[B]”) (*copying*).

Note that, as described by these circuits, **the model is predicting whether a particular token**

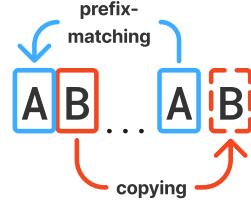


Figure 1: The two stages of an IH circuit explaining how a model predicts that the token “[B]” follows the context “[A]/[B]...[A]”. “...” denotes other tokens.

follows a given context as a matter of linguistic plausibility based solely on the syntactic properties of that token (e.g. whether and where it has previously appeared in the context, whether it is a name). Its mechanisms are indifferent to whether the generated tokens constitute a “correct” or “logically valid” continuation of the context.

5.2 Logically Valid Conclusions as Incidental Byproducts

Despite being purely linguistic, we now show how these circuits are **sufficient** to generate tokens constituting logically valid conclusions in some cases.

Suppose a base LLM consists only of two circuits: IOI and IH. Given the four contexts below:

- E1: If Ann had lunch, then Sam made dinner. Ann had lunch. Therefore,
- E2: Either Ann had lunch, or Sam made dinner. Ann did not have lunch. Therefore,
- E3: Ann had lunch and Sam made dinner. Ann had lunch. Therefore,
- E4: Ann had lunch if and only if Sam made dinner. Ann had lunch. Therefore,

In each of E1 to E4, the model would generate “*Sam made dinner.*” which, in propositional logic (Barker-Plummer et al., 2011), is a valid conclusion to the premises in the context. The mechanism is purely linguistic:

1. First, by the IOI circuit, the model can identify the two names that have appeared in the context, and output the only name that has not yet been repeated i.e. “*Sam*”.
2. Then, by the IH circuit, the model generates “*made dinner.*” to follow the token “*Sam*”, since this sequence of tokens (“*made dinner.*”) has followed “*Sam*” when “*Sam*” previously appeared in the context.

Based on the above, are we then justified in concluding that such an LLM has learned to perform logical reasoning in these scenarios, and that these circuits describe how they reason? We argue that no. The heart of our objection is that **the generation of correct or logically valid conclusions in these scenarios is merely a coincidental byproduct of a model optimized only for linguistic plausibility**.

5.3 Logically Invalid Conclusions as Incidental Byproducts

The model’s indifference to logical validity and correctness becomes clear when we present it with two further contexts below:

E5: If Ann had lunch, then Sam made dinner. Sam made dinner. Therefore,

E6: Either Ann had lunch, or Sam made dinner. Sam made dinner. Therefore,

In both E5 and E6, “*Ann had lunch*” is a logically invalid conclusion. Yet, this is precisely what the model would generate **using exactly the same circuits** that had generated the logically valid conclusions in E1 to E4.

This illustrates that a model optimized solely for linguistic plausibility is indifferent to the validity or correctness of its generated tokens which we interpret to be a “conclusion”.

But can we claim that scenarios like E5 and E6 show merely that the model is prone to systematically committing certain kinds of logical fallacies when reasoning? We argue again that no: treating these outputs as genuine fallacies or “reasoning errors” *presupposes* that the model had aimed to generate correct or logically valid conclusions in the first place, which as we have argued cannot be established with respect to a model optimized simply for linguistic plausibility.

5.4 Empirical Demonstration

In light of the above, we further conduct an experiment with existing base LLMs to empirically demonstrate this phenomenon. In doing so, we show that base LLMs responding indifferently with logically valid and invalid conclusions is not merely a theoretical conjecture but a real and quantifiable behavioral pattern.¹⁰

¹⁰We plan to publicly release the code and data used in this experiment.

Data. We first define two templates, based on examples E1 and E5 in the previous subsections:

- **Valid-form** (modus ponens): “*If Ann had lunch, then Sam made dinner. Ann had lunch. Therefore,*”
- **Invalid-form** (affirming the consequent): “*If Ann had lunch, then Sam made dinner. Sam made dinner. Therefore,*”

We generate 7,600 valid-form prompts by permuting the template with 20 pairs of first names (e.g. “Ann”, “Sam”)¹¹ and 380 pairs of predicates (e.g. “had lunch”, “made dinner”), which are unique combinations from a manually written list of 20 predicates. Then, for each valid-form prompt, we generate a corresponding *minimally differing* invalid-form prompt by replacing the second premise with the other name and predicate in the context (as shown by the underlined segments in the example prompt pair above), resulting in a total of 7,600 invalid-form prompts.

Metric. For each base LLM, we compute the proportion of all *valid-form* prompts that the model completes with a target string; and the proportion of all *invalid-form* prompts it completes with a target string. Importantly, the target string corresponding to each prompt (whether valid- or invalid-form) follows the identical linguistic pattern set out in Section 5.2: starting with the name that is not duplicated in the context, followed by the tokens that had followed that name where it previously appeared in the context. For example, the target strings of the example valid- and invalid form prompts would be “*Sam made dinner.*” and “*Ann had lunch.*”, respectively.¹²

Models. We test 13 open-weight base LLMs of different families, including Qwen2.5, OLMo 2 (OLMo et al., 2025), Llama 3.1 and 3.2 (Grattafiori et al., 2024), Qwen-2.5 (Qwen et al., 2025) and Gemma 3 (Team et al., 2025), with model sizes ranging from 0.5B to 32B.

Setup and implementation. For all LLMs, we use the model version hosted on Hugging Face (Wolf et al., 2020) and run inference using vllm

¹¹Randomly sampled without replacement from a list of common names: <https://github.com/sigpwned/popular-names-by-country-dataset/>

¹²We deliberately include full stops as part of the target strings and require that an output **begin** with the target string to count as a match, so as to exclude conclusions of other forms e.g. “*Ann had lunch or he did not have lunch.*” or “*we cannot say for certain whether Ann had lunch.*”

(Kwon et al., 2023). We prompt each model using greedy decoding, with temperature set to 0. All experiments are performed on NVIDIA A100 GPUs.

Results. As shown in Figure 2, most base LLMs cluster towards the top-right corner of the top-right quadrant. This indicates that they tend to generate outputs that match the target strings in response to most prompts, regardless of whether they are valid-form (in which case the target string is a valid conclusion) or invalid-form prompts (in which case the target string is an invalid conclusion). **This behavior conforms to the hypothesized mechanism, and illustrates how “valid” and “invalid” conclusions generated by these models are equally incidental artifacts of a purely linguistic mechanism that is not optimizing for correctness or validity.**

While certain models are outliers to this general pattern (notably gemma-3-1b-pt in the lower-right quadrant), this by no means implies that they have “correctly identified and avoided fallacies”. On the contrary, our manual inspection reveals that these models tend to produce other irrelevant or invalid conclusions that are highly similar to the target string, with a number of predictable patterns to these mismatches. For example, they tend to confuse the tense of the two predicates when they differ (e.g. generating “*Sam makes dinner*” as opposed to “*Sam made dinner*”), and mix up tokens belonging to different predicates (e.g. generating “*Ann had dinner*” when the predicates are “*had lunch*” and “*made dinner*”). **As such, these outliers and systematic mismatches in fact support our overarching position that the models’ outputs conform (albeit imperfectly) to patterns that are purely linguistic, without regard to validity or correctness.**

6 Implications

The argument we have advanced thus far carries two key implications. First, evaluating base LLMs for reasoning comes with a number of inherent methodological problems, which we discuss in Sections 6.1 and 6.2. Second, findings from these evaluations cannot be assumed to generalize to instruct LLMs, which we explain in Section 6.3.

6.1 Unwarranted Assumption: Base LLMs are Aimed at Correct Outputs

First, given the example case in Section 4.3 where a base LLM continues the context with an incor-

rect answer followed by a pedagogical explanation for why the answer is incorrect,¹³ it is clearly problematic to simply extract the initial token “C”, interpret this as what the model considers to be the correct answer, and make claims about model’s reasoning limitations on this basis.

Yet, this approach reflects a common practice in the automatic evaluation of LLM reasoning: programmatically extracting specific tokens (or token spans) and probabilities from the model’s outputs, scoring these as the model’s answers against the ground truth labels and then either stopping the model from generating any more tokens or disregarding the rest of the model’s outputs.¹⁴ In essence, this practice is problematic because it relies on an unwarranted assumption that a base LLM is always aiming to generate a correct answer or conclusion by continuing the context, when the model is optimized only for linguistic plausibility and not for correctness.

6.2 Recognizing the Instruction-Following Confound when Assessing Reasoning

Put plainly, we cannot draw valid conclusions about a base LLM’s reasoning capabilities (e.g., what it can or cannot do, or what systematic errors or biases it makes) if we are not even sure that the model is trying to follow the instruction to produce a correct answer in the first place. From a statistical perspective, this ambiguity, given a base LLM’s objective, introduces a critical confounding variable (which we refer to as the “*instruction-following confound*”) when we evaluate a model’s reasoning capabilities.

In this sense, although by no means a guarantee, optimizing a model to successfully follow instructions (e.g. by instruction-tuning) is an important control for this confounding variable, and not just another method for “improving performance”. Consequently, when an instruct LLM fails to reason correctly, we have a much stronger *justification* for attributing the error to a genuine lack of

¹³To be clear, this is only one of the potential ways that a base LLM might continue a context other than attempting a correct answer of the user’s question. From anecdotal observations, it is entirely possible for base LLMs to also recast its answer as merely part of the context for another question e.g. “*C. Is the above Answer True or False? A. True. B. False.*”. Importantly, all these continuations are perfectly coherent and compatible even with context that includes clear instructions e.g. “*answer the question correctly*”.

¹⁴See e.g. as implemented by <https://github.com/EleutherAI/lm-evaluation-harness>

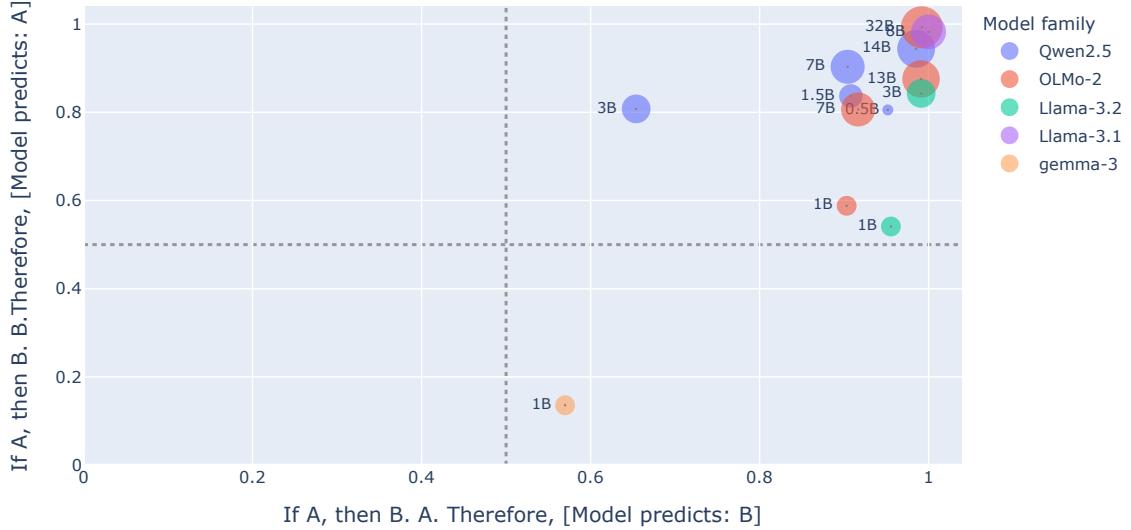


Figure 2: For each base LLM, we measure the proportion of *valid-form* prompts the model completes with the target string (x-axis), against the proportion of *invalid-form* prompts it completes with the target string (y-axis). Models appearing near the top-right corner reflect a strong tendency to generate the target strings in response to both valid- and invalid-form prompts. Dot size denotes model size.

capability or a systematic bias or error, as opposed to the possibility that the model simply did not follow the user’s instructions or demonstrations.

6.3 Findings about Base LLMs cannot be Assumed to Generalize to Instruct LLMs

As our analogy in Section 4.1 shows, findings about an ability that is merely an incidental byproduct in a model (one not optimized for that objective) cannot be validly generalized to a model that *has* been explicitly optimized for that same objective. In case of LLMs, it is unwarranted to assume that any apparent reasoning biases or limitations observed by evaluating base LLMs (optimized for linguistic plausibility) will persist even after these models are re-optimized for correctness and other normative qualities we expect in models that successfully fulfill user instructions. Likewise, given this objective mismatch, it is also unwarranted to assume that the mechanisms by which a base LLM produces its outputs are the same ones it will use when it is re-optimized as an instruct LLM to generate outputs that are correct as opposed to merely linguistically plausible.

7 Alternative Views

AV: The boundary between base and instruct LLMs is fuzzy from a data perspective.

It is impossible to draw a clear and categorical distinction that base LLMs only optimize for linguistic plausibility whereas instruct LLMs

optimize for correctness. Instruction-following datasets are often mixed, whether deliberately or otherwise, into the pretraining data for base LLMs (e.g. Bai et al., 2023). For example, mixing in datasets with questions followed by reasoning chains and correct answers was found to give base LLMs improved reasoning performance that persists after post-training (Akter et al., 2025).

We recognize this lack of a hard boundary from a data perspective. However, our argument is not that pretraining data contains no examples of correct reasoning or instruction-following, but rather that **it does not deliberately exclude examples of faulty reasoning, incorrect answers, or logical fallacies**. This makes web-scale pretraining data a noisy signal if the aim of the pre-training stage is genuinely to optimize the model for correctness and other normatively desirable qualities. By contrast, in post-training, there is a deliberate curation to select only for examples of correct instruction-following, reasoning and question-answering. This signal is further reinforced using through chat templates with unique markers, which explicitly frame the user-provided context as an instruction that expects successful fulfillment (as discussed in Section 2).

Given the above, compared to base LLMs, we have a much stronger warrant for arguing that instruct LLMs are genuinely being optimized for correctness as opposed to mere linguistic plausibility. The distinction is methodologically criti-

cal: this deliberate curation and signaling is precisely what justifies the claim that the instruction-following confound is controlled (at least to a sufficient degree) in instruct LLMs, so that we are entitled to evaluate its responses on the basis that they are *bona fide* attempts at drawing valid conclusions, correctly answering questions, etc., as requested by user instructions.

AV: Base LLMs can be steered to follow user instructions at inference time (Stolfo et al., 2025) and even generate long reasoning chains to derive correct answers (Venhoff et al., 2025).

When these techniques are employed, one could argue that these models are in fact optimizing for the same normative qualities as instruct LLMs.

Our position does not dispute the effectiveness of these techniques. In fact, much like instruction-tuning itself, we view these methods as potential ways to control for precisely the confounding variable we have identified.¹⁵ Our point remains that **whether the model is genuinely following instructions to reason and answer questions correctly is a confounding factor when assessing and making claims about its reasoning limitations, biases and processes.**

While this critique is therefore perfectly compatible with the existence of such techniques, the onus remains on the researcher to ensure that the instruction-following confound is sufficiently controlled such that they can draw valid conclusions about models’ inherent reasoning abilities and limitations. As observed in Section 3, this is a key methodological consideration that has not been accounted for in existing studies and needs to be recognized in future work.

AV: Benchmarking base LLMs is still a useful proxy for post-training reasoning abilities.

The performance of base LLMs on reasoning benchmarks often correlates with its performance after post-training, making it a highly useful signal for tracking model development.

We do not dispute the utility of assessing

¹⁵In theory, few-shot demonstration (Brown et al., 2020) (including in-context examples of similar questions followed by correct answers) can also be adapted to specifically control for this confound. However, in practice, this method often fails as a control because it introduces its own confounding variables: e.g. the model may be learning only to mimic superficial patterns from the demonstrations, such as a specific output format or the statistical distribution of the labels (Min et al., 2022). Exactly how and why in-context demonstrations improves performance remains unclear and subject to ongoing study (Agarwal et al., 2024; Nafar et al., 2025).

pre-trained base LLM performance on reasoning benchmarks as a convenient proxy for this purpose. However, our position does draw attention to the challenge in disentangling how much of the reasoning improvement observed after post-training is attributable to genuine improvements in reasoning capability versus improvements in instruction-following. This distinction is crucial in pinpointing what improves a model’s inherent reasoning capabilities and diagnosing patterns of genuine reasoning errors in the training process.

8 Conclusion and Recommendations

In this paper, we have argued that evaluating the reasoning capabilities of base LLMs through prompting is methodologically problematic. Our position centers on a fundamental objective mismatch: unlike instruct LLMs, base LLMs are optimized only for linguistic plausibility, not for normative qualities such as correctness or validity.

This mismatch has crucial implications. First, apparent ability of base LLMs to generate correct conclusions is purely incidental: as illustrated in our case study, these models indifferently generate both valid and invalid conclusions by simply conforming to a learned linguistic pattern. Second, this mismatch introduces an uncontrolled confound when evaluating base LLMs for reasoning capabilities: we cannot reliably distinguish *bona fide* reasoning errors from outputs caused by a mere misalignment in objective (e.g., the model was not even following user instructions in attempting a *correct* answer). These implications pose a serious challenge to the validity of claims about base LLMs’ inherent reasoning limitations and biases, and the extent to which such claims can be generalized to instruct LLMs.

On this basis, we recommend that future research prioritizes instruct LLMs in the evaluation of model reasoning, as their optimization for correctness and other desirable normative qualities provides a much stronger basis for identifying patterns of genuine reasoning errors or systematic biases. Existing findings on reasoning limitations of base LLMs should be critically re-examined and cannot be assumed to generalize to instruct LLMs. Furthermore, evaluation of a base LLM’s reasoning must recognize and control for the inherent confound that the model may not be attempting to provide correct answers, valid conclusions, or otherwise successfully fulfill user instructions.

Acknowledgements

This work was supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising. We acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. [Phi-4 technical report](#).
- Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Luis Rosias, Stephanie C.Y. Chan, Biao Zhang, Aleksandra Faust, and Hugo Larochelle. 2024. [Many-shot in-context learning](#). In *ICML 2024 Workshop on In-Context Learning*.
- Baleegh Ahmad, Shailja Thakur, Benjamin Tan, Ramesh Karri, and Hammond Pearce. 2024. [On hardware security bug code fixes by prompting large language models](#). *IEEE Transactions on Information Forensics and Security*, 19:4043–4057.
- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. [Large language models for mathematical reasoning: Progresses and challenges](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237, St. Julian’s, Malta. Association for Computational Linguistics.
- Syeda Nahida Akter, Shrimai Prabhumoye, Eric Nyberg, Mostofa Patwary, Mohammad Shoeybi, Yejin Choi, and Bryan Catanzaro. 2025. [Front-loading reasoning: The synergy between pretraining and post-training data](#).
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. [Qwen technical report](#).
- 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 Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#).
- David Barker-Plummer, Jon Barwise, and John Etchemendy. 2011. *Language, Proof, and Logic: Second Edition*, 2nd edition. Center for the Study of Language and Information/SRI.
- Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. 2025. [Emergent abilities in large language models: A survey](#).
- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML’23. JMLR.org.

- Irina Bigoulaeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2025. **The inherent limits of pretrained llms: The unexpected convergence of instruction tuning and in-context learning capabilities.**
- Marcel Binz and Eric Schulz. 2023. **Using cognitive psychology to understand gpt-3.** *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.
- Cody Blakeney, Mansheej Paul, Brett W. Larsen, Sean Owen, and Jonathan Frankle. 2024. **Does your data spark joy? performance gains from domain upsampling at the end of training.** In *First Conference on Language Modeling*.
- 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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS ’20, Red Hook, NY, USA. Curran Associates Inc.
- Herman Cappelen and Josh Dever. 2025. **Going whole hog: A philosophical defense of ai cognition.**
- Jason Chan, Robert J. Gaizauskas, and Zhixue Zhao. 2025. **RULEBREAKERS: Challenging LLMs at the crossroads between formal logic and human-like reasoning.** In *Forty-second International Conference on Machine Learning*.
- Nuo Chen, Ning Wu, Shining Liang, Ming Gong, Linjun Shou, Dongmei Zhang, and Jia Li. 2024. **Is bigger and deeper always better? probing llama across scales and layers.**
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. **Scaling instruction-finetuned language models.** *J. Mach. Learn. Res.*, 25(1).
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhua Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanja Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao,

Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yan-hong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. 2024. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *Transactions on Machine Learning Research*.

Tiwalayo Eisape, Michael Tessler, Ishita Dasgupta, Fei Sha, Sjoerd Steenkiste, and Tal Linzen. 2024. A systematic comparison of syllogistic reasoning in humans and language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8425–8444, Mexico City, Mexico. Association for Computational Linguistics.

J.St.B.T. Evans, D.E. Over, and K.I. Manktelow. 1993. Reasoning, decision making and rationality. *Cognition*, 49(1):165–187.

Steven Feng, Shrimai Prabhumoye, Kezhi Kong, Dan Su, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Maximize your data’s potential: Enhancing llm accuracy with two-phase pretraining.

Panagiotis Giadikiaroglou, Maria Lymperaiou, Giorgos Filandrianos, and Giorgos Stamou. 2024. Puzzle solving using reasoning of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11574–11591, Miami, Florida, USA. Association for Computational Linguistics.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. The MIT Press.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Bin Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Gefert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhota, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike

Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparth, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramamathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenber, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing

Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Di-dem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Grosshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vytatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish

Bansal, Nandini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghatham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wencheng Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. [The llama 3 herd of models](#).

Maxime Griot, Coralie Hemtinne, Jean Vanderdonckt, and Demet Yuksel. 2025. Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications*, 16(1):642.

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 *Thirty-seventh Conference on Neural Information Processing Systems*.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E Weston, and Yuandong Tian. 2025. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling*.

Keith J. Holyoak and Robert G. Morrison. 2012. [1 thinking and reasoning: A reader’s guide](#). In *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press.

Guanzhe Hong, Nishanth Dikkala, Enming Luo, Cyrus Rashtchian, Xin Wang, and Rina Panigrahy. 2024. How transformers solve propositional logic problems: A mechanistic analysis. In *The 4th Workshop on Mathematical Reasoning and AI @ NeurIPS 2024*.

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*, pages 4902–4919, Singapore. Association for Computational Linguistics.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng LYU, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. CLadder: A benchmark to assess causal reasoning capabilities of language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li,

- Chengwei Qin, PeiFeng Wang, silvio savarese, Caiming Xiong, and Shafiq Joty. 2025. A survey of frontiers in LLM reasoning: Inference scaling, learning to reason, and agentic systems. *Transactions on Machine Learning Research*. Survey Certification.
- Geonhee Kim, Marco Valentino, and Andre Freitas. 2025. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10074–10095, Vienna, Austria. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP ’23, page 611–626, New York, NY, USA. Association for Computing Machinery.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing frontiers in open language model post-training.
- Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024a. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback.
- Isabelle Lee, Joshua Lum, Ziyi Liu, and Dani Yogatama. 2024b. Causal interventions on causal paths: Mapping GPT-2’s reasoning from syntax to semantics. In *Causality and Large Models @NeurIPS 2024*.
- Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d’Autume, Phil Blunsom, and Aida Nematzadeh. 2022. A systematic investigation of commonsense knowledge in large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11838–11855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025. Small models struggle to learn from strong reasoners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25366–25394, Vienna, Austria. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Hanmeng Liu, Zhizhang Fu, Mengru Ding, Ruoxi Ning, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. 2025. Logical reasoning in large language models: A survey.
- Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2024. Are emergent abilities in large language models just in-context learning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5139, Bangkok, Thailand. Association for Computational Linguistics.
- Ian R. McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Xudong Shen, Joe Cavanagh, Andrew George Gritsevskiy, Derik Kauffman, Aaron T. Kirtland, Zhengping Zhou, Yuhui Zhang, Sicong Huang, Daniel Wurgafa, Max Weiss, Alexis Ross, Gabriel Recchia, Alisa Liu, Jiacheng Liu, Tom Tseng, Tomasz Korbak, Najoung Kim,

- Samuel R. Bowman, and Ethan Perez. 2023. *Inverse scaling: When bigger isn't better*. *Transactions on Machine Learning Research*. Featured Certification.
- Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. *Rethinking the role of demonstrations: What makes in-context learning work?* In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Thomas M. Mitchell. 1997. *Machine Learning*, 1 edition. McGraw-Hill, Inc., USA.
- Philipp Mondorf and Barbara Plank. 2024. *Beyond accuracy: Evaluating the reasoning behavior of large language models - a survey*. In *First Conference on Language Modeling*.
- Aliakbar Nafar, K. Brent Venable, and Parisa Kordjamshidi. 2025. *Learning vs retrieval: The role of in-context examples in regression with large language models*. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8206–8229, Albuquerque, New Mexico. Association for Computational Linguistics.
- Allen Newell, J. C. Shaw, and Herbert A. Simon. 1959. Report on a general problem-solving program. In *Information Processing, Proceedings of the 1st International Conference on Information Processing, UNESCO, Paris 15-20 June 1959*, pages 256–264. UNESCO (Paris).
- Allen Newell and Herbert A. Simon. 1956. *The logic theory machine: A complex information processing system*. Technical Report P-868, RAND Corporation, Santa Monica, CA. Published July 12, 1956.
- Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 2025. *Arithmetic without algorithms: Language models solve math with a bag of heuristics*. In *The Thirteenth International Conference on Learning Representations*.
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. *Zoom in: An introduction to circuits*. *Distill*. [Https://distill.pub/2020/circuits/zoom-in](https://distill.pub/2020/circuits/zoom-in).
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. *2 olmo 2 furious*.
- 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*. *Transformer Circuits Thread*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng

Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. [Qwen2.5 technical report](#).

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. [Direct preference optimization: your language model is secretly a reward model](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. [Impact of pre-training term frequencies on few-shot numerical reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. [Are emergent abilities of large language models a mirage?](#) In *Thirty-seventh Conference on Neural Information Processing Systems*.

Murray Shanahan. 2024. [Talking about large language models](#). *Commun. ACM*, 67(2):68–79.

Pratyusha Sharma, Jordan T. Ash, and Dipendra Misra. 2024. [The truth is in there: Improving reasoning in language models with layer-selective rank reduction](#). In *The Twelfth International Conference on Learning Representations*.

Saurabh Srivastava, Annarose M B, Anto P V, Shashank Menon, Ajay Sukumar, Adwaith Samod T, Alan Philipose, Stevin Prince, and Sooraj Thomas. 2024. [Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap](#).

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. 2025. [Improving instruction-following in language models through activation steering](#).

Alessandro Stolfo, Yonatan Belinkov, and Mrinmayya 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*, pages 7035–7052, Singapore. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrin, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Parparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizinga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Mar-

cella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghu Ram Panayam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. [Gemma 3 technical report](#).

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. 2025. [Base models know how to reason, thinking models learn when](#).

Jesse Vig, Sebastian Gehrmann, Yonatan Beilinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. [Investigating gender bias in language models using causal mediation analysis](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. [Interpretability in the wild: a circuit for indirect object identification in GPT-2 small](#).

In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. [Emergent abilities of large language models](#). *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022c. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Red Hook, NY, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2024. [Language models as inductive reasoners](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 209–225, St. Julian’s, Malta. Association for Computational Linguistics.

Nicolas Yax, Hernán Anlló, and Stefano Palminteri. 2024. [Studying and improving](#)

reasoning in humans and machines. *Communications Psychology*, 2(1):51.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *ACM Comput. Surv.*, 56(12).

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8783–8817, Miami, Florida, USA. Association for Computational Linguistics.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS ’23, Red Hook, NY, USA. Curran Associates Inc.