Towards *Computational Comprehension*: A Non-Anthropocentric Framework for Evaluating LLM Understanding

Anonymous ACL submission

Abstract

This position paper introduces and motivates **Computational Comprehension** as an alternative, non-anthropocentric approach to assessing how Large Language Models (LLMs) handle knowledge. Unlike standard benchmarks, which often reward models for surface-level accuracy but shed little light on deeper conceptual understanding, Computational Comprehension directs attention to the model's internal processes. Specifically, we focus on whether certain neurons or sub-networks remain consistently activated across reformulations and contextual shifts of the same underlying concept. We outline a framework that tests a model's ability to preserve conceptual invariance under various input transformations, then observes how targeted ablations of relevant sub-networks affect performance. By gauging these internal, concept-related responses rather than relying solely on external metrics, we obtain more fine-grained insights into a model's capacity to internalize, manipulate, and robustly apply conceptual knowledge. We also propose integrating such analysis into systematic experiments, showing how subtle tweaks to task prompts or data can reveal whether a model is genuinely *concept-driven* or merely *parroting surface correlations*. Through Computational Comprehension, we encourage researchers, engineers, and theorists to adopt a deeper, more transparent mode of evaluation—one that foregrounds *internal conceptual grounding* over score-centric arms races in pursuit of ever-higher benchmark numbers.

1 Introduction

Designing and evaluating Large Language Models (LLMs) has long relied on human-centric notions of *understanding*, shaped by both philosophical definitions and our intuitive experience as humans. Traditionally, the term *understanding* invokes a cognitive process in which a person internalizes and applies concepts to model an event, context, or a piece of information (Bereiter, 2002). However, this anthropocentric framework creates immediate challenges for modern AI systems: ascribing human-style mental states or semantic understanding to models poses possibility of conflating symbol manipulation with genuine comprehension (Bennett, 2023). Recent arguments—including the Chinese Room thought experiment (Searle, 1980), the Mary's Room scenario (Jackson, 1986), and the Stochastic Parrot critique (Bender et al., 2021)—have raised doubts as to whether text-based LLMs can possess any more than surface-level patterns. Indeed, such arguments underscore a deeper question: *in what sense can an artificial system be said to understand at all?*

We propose a new term, **Computational Comprehension**, to distinguish the kind of understanding that may be attained by LLMs and other advanced neural architectures. Unlike human cognition, which is heavily shaped by embodiment, sensory modalities, and evolution (Foglia and Wilson, 2013), Computational Comprehension emphasizes the unique capacity of a model to form, store, and manipulate its own internal representations (or features) in a manner that enables consistent generalization and coherent behavior.

In doing so, we acknowledge both the findings of *mechanistic interpretability* research—which highlights emergent circuits that correlate with concept-like features (Olah et al., 2020)—and the doubt that arises when we observe these models hallucinating, making errors, or merely pattern-matching pre-trained data. By framing an alternative, *non-anthropocentric* construct, our notion of Computational Comprehension aims to capture both the potential and the limitations of such systems without presupposing human-like consciousness or qualia.

A critical motivation for introducing and clarifying this term is the evolving crisis in LLM benchmarks. Many of today's LLMs not only surpass the

average human in certain standardized tests but are rapidly approaching near-perfect performance on tasks designed by humans (Strachan et al., 2024). As a consequence, human score on benchmarks no longer provides a sufficient yardstick (Hern´andez-Orallo, 2020). If two models both exceed human excellence (near-100% accuracy), how should one meaningfully evaluate their capabilities? We argue that simply escalating benchmark difficulty is not a robust long-term solution. Instead, *we must adopt an orthogonal, conceptual axis of evaluation: a model's degree of Computational Comprehension*.

By investigating whether models truly capture, adapt, and integrate their internal features—as opposed to *merely* reproducing text patterns reviewed in training, we gain a more nuanced, explanatory picture of how they handle knowledge. Examining Computational Comprehension thus becomes a crucial research direction: it reframes the debate on LLMs' understanding, guides interpretability studies, and opens new vistas for building an evaluation framework that reflect real conceptual robustness.

In this paper, we elaborate upon the philosophical underpinnings that motivate this new terminology and *propose initial pathways to measure Computational Comprehension in practice*. By doing so, we seek to foster a framework that meaningfully captures the ways in which next-generation models process *concepts*, even if that process differs radically from human cognition. Our goal is not merely to defend a novel philosophical position but to engage practitioners and theorists alike in rethinking how we might define, test, and refine model-based comprehension without anchoring it to a solely anthropocentric conceptual lineage. Throughout this paper, we speak of *model understanding* and Computational Comprehension in the specific context of text-based LLMs.

2 Philosophical & Cognitive Foundations

In this section, we lay out the philosophical and cognitive cornerstones that shape our understanding of *understanding* itself. By examining long-standing debates about human cognition, we highlight the anthropocentric assumptions often embedded in discussions of AI (Millière and Rathkopf, 2024). We begin by introducing how the human basis of understanding informs (and potentially biases) our interpretation of large language model capabilities, then turn to three well-known arguments that each challenge whether symbolic proficiency can ever equate to genuine comprehension.

2.1 The Human Basis of Understanding

The concept of *understanding* has been explored by philosophers and cognitive scientists for centuries, commonly framed as a process wherein an individual internalizes and applies concepts to interpret or predict phenomena in the world (Bereiter, 2002). However, these perspectives often adopt a human-centric perspective, or anthropocentrism, which positions human beings at the center of consideration. In this viewpoint, everything from philosophical inquiries to recent developments, such as LLMs, are often evaluated through human standards, reflecting the belief that human knowledge, reasoning, and consciousness are the ultimate criteria by which we measure understanding.

Yet the question remains whether this anthropocentric concepts translates readily to artificial architectures. These architectures may excel at pattern recognition without replicating *human subjective dimensions* (consciousness, experiences, etc.), which philosophers often regard essential to genuine comprehension (Foglia and Wilson, 2013). This raises the question *Can LLMs' internal parameters actually constitute features analogous to human concepts, or are they simply finely tuned statistical representations?*

2.2 Central Challenges: Three Arguments

To answer this question, we begin by examining *Searle's Chinese Room* (Searle, 1980), *Mary's Room* (Jackson, 1986), and the *Stochastic Parrot* critique (Bender et al., 2021), each aiming to highlight a distinct aspect of the gap between symbolic prowess and what human-centric notions would consider true comprehension.

Searle's Chinese Room

John Searle's Chinese Room thought experiment argues that the manipulation of symbols according to formal rules—yielding outputs indistinguishable from those of a fluent speaker—does not constitute genuine understanding. In this scenario, a person who speaks only English follows instructions to map Chinese input symbols to output symbols, appearing to understand Chinese, yet lacks any real semantic comprehension. This mirrors concerns in AI that models may generate appropriate responses without internal comprehension, effectively acting as sophisticated rulebooks where substantial parameter weights guide permissible patterns with-

out innate semantic grounding. Some argue that modern neural networks might develop emergent properties transcending simple syntactic manipulation (Wei et al., 2022), but whether these suffice for true semantic understanding remains highly debated, making Searle's thought experiment crucial to current discussions about AI cognition.

Mary's Room

In Frank Jackson's narrative, Mary is a color scientist with complete theoretical knowledge about color perception, yet she has only experienced a grayscale world. Upon actually *seeing* red, she gains new insights—qualitative knowledge about color that no enumeration of facts could convey. Initially challenging materialism in the philosophy of mind, this scenario also highlights the significance of subjective experience, beyond mere propositional facts, in understanding. For AI, the Mary's Room thought experiment underscores how purely symbolic or descriptive mastery might fail to capture certain dimensions of understanding, particularly those reliant on firsthand, qualitative experience. Since most LLMs are trained on linguistic data alone (Collier et al., 2022), critics argue these models are inherently limited to factual or correlational knowledge without the capacity for perceptual immersion. For those who see a tight bond between experience and comprehension, the thought experiment provides a powerful rejoinder to claims that advanced text models truly *know* or *understand* phenomena like color.

Stochastic Parrot Critique

Bender and colleagues (Bender et al., 2021) argue that LLMs generate text similar to meaningful human discourse but do so primarily through probabilistic rearrangement of learned patterns. These systems, as a result, assemble outputs in a manner that exhibits impressive fluency without necessarily having any underlying conceptual or intentional structure. From this perspective, an AI's *knowledge* is a patchwork drawn from extensive training data, appearing coherent because it overlays existing linguistic tropes rather than genuinely grasping the subject matter. This critique warns against attributing true *understanding* simply because a machine's output aligns with human expectations, especially as LLMs excel at tasks traditionally associated with human cognition. Proponents argue that sufficiently complex systems might develop emergent properties beyond mere pattern manipulation (Wei et al., 2022). Nonetheless, we face the dilemma of discerning whether we are witnessing a qualitatively novel form of cognition or merely the expansion of surface-level pattern recognition.

3 Hints of Artificial Comprehension

Mechanistic Interpretability

Mechanistic interpretability aims to open the *black box* of large-scale neural networks by uncovering how internal components—layers, neurons, attention heads, or circuits—instantiate concept-specific computations (Bereska and Gavves, 2024). Various interpretability techniques reveal a model's internal "concept-like" structures, or features (Templeton et al., 2024; Gao et al., 2024). Circuit tracing with probing classifiers identifies neurons that consistently fire for certain inputs (Olah et al., 2020), while feature visualization and activation patching highlight the tokens or activations most responsible for concept processing (Heimersheim and Nanda, 2024; Meng et al., 2022). Ablation studies further show that disabling suspect neurons degrades performance on targeted tasks (Yu and Ananiadou, 2024b; Zhang et al., 2024; Yu and Ananiadou, 2024a). By these means, we can often pinpoint how a model implicitly organizes its representation of knowledge. Empirical work demonstrates that removing these *concept-circuits* collapses functionality (Olah et al., 2020), suggesting dedicated internal pathways for distinct functions.

Emergent Phenomena

Grokking (Power et al., 2022) and prompt engineering such as Pause Tokens (Goyal et al., 2023) both exemplify intriguing emergent phenomena that hint at latent conceptual capacities. In *grokking*, models may plateau for many epochs before abruptly converging to near-perfect performance, suggesting a phase transition from rote memorization to a more generalizable, concept-based representation. Likewise, inserting seemingly trivial instructions or *pause tokens* can dramatically alter a model's output, apparently switching on previously dormant internal circuits. Mechanistic analyses of these effects indicate that large language models can harbor latent conceptual structures (Nanda et al., 2023), which remain quiescent until triggered by just the right input cues.

Next Step: *Computational Comprehension*

Mechanistic interpretability and emergent phenomena each shed light on a model's capacity for true Computational Comprehension. From circuit tracing and ablation studies, we learn how knowledge may be internally structured: identifying subnetworks that directly handle abstract concepts suggests more than mere pattern memorization. Likewise, phenomena like grokking and the dramatic effects of pause tokens highlight a network's latent dispositions, revealing how a seemingly small change in prompts or training conditions can trigger a deeper, more coherent conceptual state. Together, these findings go beyond measuring whether a model performs well on benchmarks: they offer insight into how robustly and systematically the model organizes its knowledge, thereby hinting at (or refuting) the presence of enduring, concept-like structures that distinguish superficial patterning from genuine comprehension.

4 Towards *Computational Comprehension*

Recent LLMs exhibit near human performance, while arguments from section 2 emphasize the inability of LLMs to acquire the experiential or conscious dimensions typically associated with genuine comprehension. We introduce Computational Comprehension as a middle ground approach: an explicitly *non*-anthropocentric notion of what it might mean for a machine to *understand* while remaining faithful to the underlying computational mechanisms and constraints.

4.1 Defining Computational Comprehension

We propose that Computational Comprehension refers to the capacity of a computational model to:

- **Internalize and represent a stable concept structure** through its internal parameters or sub-networks, rather than mere memorization of specific input–output pairs.
- **Exhibit systematic responses across diverse contexts** that test multiple dimensions of a given concept, demonstrating an ability to generalize.
- **Manipulate these internal concept-like representations** in a way that allows for novel combinations, inferences, or transformation that align with the concept's formal or functional properties.

4.2 Key Dimensions of Computational Comprehension

Alongside the core definition, we propose several dimensions that can help refine how one might observe or measure Computational Comprehension in practice:

Generality & Robustness Possessing generality and being robust allows a model to effectively apply its understanding of concepts across various tasks and contexts, demonstrating that its knowledge goes beyond memorization. For example, a system trained for addition should correctly solve novel sums presented in different formats, including newly structured inputs or unfamiliar story-based settings, and it should maintain accuracy when problems are paraphrased or slightly altered, rather than failing due to superficial modifications. Together, these attributes emphasize the depth of conceptual understanding necessary for effective machine comprehension.

Sub-Network Localization and Interpretability Identifying neurons, attention heads, or circuits that consistently activate for a specific concept offers deeper insight into whether the concept is truly internalized. Coherent sub-networks that prove indispensable across tasks relating to the concept point to a stable, localized representation.

Compositional Integration If the model can manipulate concept A and concept B, does it also reliably handle combinations of A and B (e.g., nested constructs, intersection of properties), a hallmark of truly concept-based reasoning?

This characterization steers away from ascribing any notion of *phenomenal consciousness* or *qualia* to the model. Instead, *comprehension* here is pinned to the *use* of internally stored features that enable conceptual coherence. By being explicit about what these features (and their use) look like in computational terms, we aim to avoid unnecessary anthropocentrism and keep the focus on *how* the LLMs manipulate internal concepts into consistent outputs.

5 Benchmarking, Limitations, and the Need for a New Paradigm

As large language models (LLMs) have rapidly improved, benchmark performance has been the primary metric for gauging progress. While advancements in tasks like translation and question answering are impressive, it remains unclear whether high

scores on these metrics indicate any deeper form of *understanding*(Mondorf and Plank, 2024), or Computational Comprehension. We examine the limitations of benchmark-driven evaluations and argue for a new paradigm, focused on concept-based perspectives.

5.1 Historical Context of Benchmarks and Accelerating Model Capabilities

Traditional benchmarks such as GLUE (Wang, 2018), SuperGLUE (Wang et al., 2019), SQuAD (Rajpurkar, 2016), MMLU (Hendrycks et al., 2020), and Big-Bench (Srivastava et al., 2022) have become standard metrics for evaluating progress in natural language processing (Fourrier et al., 2024), while modern LLMs achieve scores that rival or surpass average human baselines. While these advancements demonstrate the models' ability to capture complex statistical regularities, a concerning pattern emerges: once a benchmark gains recognition, models are optimized to reach near the theoretical ceiling (Zhou et al., 2023). This leads us to question what these scores truly signify: do they reflect a meaningful *understanding* of the underlying principles, or are they merely the result of memorizing patterns and exploiting correlations?

5.2 Why High Benchmark Scores May Not Imply Real Comprehension

A major critique of high benchmark scores is that models can achieve strong performances through superficial pattern matching and memorization rather than true understanding (Bender et al., 2021). For instance, when a dataset establishes consistent associations between specific words and outcomes, a model may learn to recognize these patterns akin to a *lookup table*, rather than forming robust, concept-level generalizations. This phenomenon can lead a model to excel on a particular dataset yet struggle with slight rephrasings or different contexts of the same task. Consequently, this raises the question of whether such high scores truly reflect Computational Comprehension.

5.3 Human-Centric Benchmarking Dilemma

Beyond Human-Level Evaluation

Many of the most widely recognized benchmarks were either explicitly designed to capture difficulties humans face or implicitly assumed that human performance provides an upper bound (Zhou et al., 2023). Now that models are meeting or exceeding human baselines, a paradox arises: if both Model A and Model B surpass average human performance, how do we evaluate which model is better in a conceptual sense? Traditional metrics like accuracy or F1 score, lose discriminative power once performance clusters near 100%.

Eschewing an Endless Benchmark Arms Race

An intuitive reaction might be to escalate benchmark difficulty (Arora et al., 2023), creating new sets of even more intricate questions that push models further. However, an endless cycle of difficult benchmark creation risks feeding back into a purely performance-driven arms race. Such an approach may continue yielding impressive numeric gains but could neglect how models handle concepts and whether they truly internalize them or simply scale up memorization. Unless we adopt new perspectives, the gap between high test accuracy and genuine conceptual capacity may widen.

5.4 Criteria for a New Paradigm

Orthogonal to Difficulty

A key realization is that we do not need to simply *increase* the difficulty of tasks. Instead, we must develop *orthogonal* criteria that evaluate how robustly and systematically a model uses internal representations. For instance, a model simultaneously achieving perfect accuracy and high *conceptual instability* under small rephrasings or domain shifts indicates a shallow conceptual grounding. Conversely, a model with more modest performance but stable concept usage across transformations—like those described in §4—might be argued to exhibit deeper Computational Comprehension (Nanda et al., 2023).

From Output Scores to Conceptual Analysis

Traditional accuracy-based benchmarks measure *what* the model outputs but not *how* it arrives at those outputs. Mechanistic interpretability (§3) opens a window into *where* concept-like circuits reside (Olah et al., 2020), and *how* the model manipulates them. By integrating such insights into a new evaluation paradigm, researchers can design tasks and metrics that test for *invariance, compositional integration,* and *generalizable conceptual structure* rather than pure output correctness (Zahraei and Asgari, 2024). This reframes model assessment as probing the alignment between *internal computations* and *expected conceptual consistency*.

5.5 How Computational Comprehension Strengthens Model Evaluation

Focusing on Computational Comprehension offers a nuanced approach to evaluating model performance beyond traditional benchmarks. For instance, when Model A and B achieve identical scores, an analysis of Computational Comprehension may show that Model A employs identifiable concept circuits with robustness to synonymous phrases and logical coherence in reasoning, while Model B falters under minor perturbations and lacks clear internal structures, indicating that Model A demonstrates superior understanding despite similar benchmark scores. This approach reduces the risk of relying on superficial performance gains, as it requires models to demonstrate stable internal feature usage across a variety of tasks.

5.6 Addressing Objections and Feasibility

Critics of concept-centric metrics argue that implementing repeated transformations and deep ablation studies can be more resource-intensive than numeric benchmarks, where a single accuracy or F1 score provides straightforward model comparison. However, this added complexity is essential for exposing deeper strengths and weaknesses in model cognition. Observational parallels can be drawn to neuroscience and psychology, fields that similarly adopt intricate testing to capture complex phenomena. Importantly, Computational Comprehension metrics do not displace existing evaluations; rather, they complement them. Conventional benchmarks remain invaluable for quickly gauging baseline capability or domain-specific performance. By layering in concept-invariance tests and interpretability measures, we can build a more complete picture of how robustly and structurally a model encodes the knowledge it appears to exhibit.

6 Formal Framework for Assessing *Computational Comprehension*

This section proposes a formal framework for assessing Computational Comprehension, uniting these ideas into a methodology that goes beyond simple output metrics. In doing so, we shift emphasis from "*How high is the accuracy?*" to "*How stable and interpretable are the model's concept representations?*"

6.1 Concepts as Functional Mappings

We formalize a *concept* as a relation (or function) that maps a set of inputs $\mathcal{X}$ to a corresponding set of outputs or labels $\mathcal{Y}$. Concretely, we write:

$$C \colon \mathcal{X} \to \mathcal{Y},$$

where $C(x) \in \mathcal{Y}$ represents the ideal conceptual outcome (or property) for each input $x \in \mathcal{X}$. Crucially, C captures the *principle* or *logic* dictating how inputs relate to outputs, rather than a mere memorized pairing. For example, $C(x)$ could designate whether x meets certain criteria, identifies a particular feature, or even transforms x according to a well-defined rule.

From this standpoint, a model M *masters* a concept C if it satisfies $M(x) \approx C(x)$ for all relevant inputs $x \in \mathcal{X}$. Such mastery demands more than just fitting a fixed set of training examples—it requires that M capture and robustly apply the underlying *principle* unifying all $(x, C(x))$ pairs. In other words, M should exhibit:

- **Consistency:** The model's predictions or outputs for new examples x' reflect the same structural or semantic rule encoded by C.
- **Generality:** The learned concept holds over the entire space $\mathcal{X}$ (or its relevant subset), not solely for the specific instances seen during training.
- **Stability:** Small perturbations to x that preserve conceptual properties should not disrupt M's alignment with C, signaling that M latently encodes the concept rather than memorizing superficial patterns.

If the model genuinely *understands* a concept, it will consistently and accurately reproduce $C(x)$ across transformations and contexts that preserve the conceptual core.

6.2 Transformational Robustness

One hallmark of true Computational Comprehension is the capacity to preserve conceptual validity under input transformations that leave the *concept* unchanged. Formally, suppose we define a set T of transformations $t : X \to X$ that maintain $C(x)$ in the sense that:

$$C\big(t(x)\big) = C(x),$$

for all $t \in T$ and $x \in X$. If a model M genuinely grasps C, it should map $t(x)$ to the same conceptual outcome as x—that is, $M(\,t(x)\,) \approx M(x)$ whenever $C(\,t(x)\,) = C(x)$.

Benchmark Extension. Instead of scoring M solely on the raw input–output pairs $(x, C(x))$, we propose also testing across transformations $(t(x), C\big(t(x)\big))$. The model's *robustness* score $R(C, M)$ measures consistency under these transformations:

$$R(C, M) \;=\; \mathbb{E}_{(x,t)\,\in\,(X\times T)}\Big[\mathbf{1}(M(t(x)) \;=\; C(t(x)))\Big]$$

High $R(C, M)$ suggests the model's internal representation is more than superficially memorizing $(x, C(x))$ pairs; it is capturing the general rule that remains invariant under t.

6.3 Linking Internal Representations via Mechanistic Interpretability

Even high robustness could, in principle, arise from sophisticated but purely *opaque* pattern-fitting. To address this risk, we invoke *mechanistic interpretability* (§3). Concretely, we hypothesize that for each concept C, there should exist a *sub-network* in M—some ensemble of neurons, attention heads, or circuits—that is critical to consistently mapping x to $C(x)$. By identifying and testing that sub-network, we can verify whether:

1. **Localization.** Activations intensify in the sub-network specifically when the model processes inputs related to C, compared to unrelated tasks.

2. **Ablation Degrades C.** Disabling or corrupting the targeted sub-network (e.g., *zero/mean ablation*) yields a sizable drop in $R(C, M)$.

3. **Consistency under Transformations.** The same sub-network remains activated (or similarly structured) even for transformed inputs $t(x)$, reinforcing that M is implementing C in a stable, generalizable manner.

These conditions connect conceptual invariance to an *observable* internal mechanism, offering a more holistic demonstration of Computational Comprehension.

6.4 Proposed Metrics and Procedures

We outline a possible experimental procedure, synthesizing ideas from previous sections:

I. **Identify a Target Concept C.** Formally define the function or relation you want to test (e.g., logical equivalence, color categorization, basic arithmetic).

II. **Generate Transformation Family T.** Construct transformations $t \in T$ that preserve C, such as paraphrases, format changes, or domain shifts.

III. **Compute Baseline Accuracy.** First assess M on pairs $(x, C(x))$ to ensure M handles straightforward versions of C acceptably.

IV. **Evaluate Robustness $R(C, M)$.** Arrange $t(x)$ for $t \in T$ on test inputs, verifying if $M(t(x)) \approx C\big(t(x)\big) = C(x)$.

V. **Mechanistic Analysis.**

   Find Relevant Circuits/Neurons. Use probing, circuit tracing, and feature visualization (see §3) to locate internal structures maintaining C.

   Perform Ablation. Temporarily ablate (e.g. zero ablation) the identified sub-network. Does $R(C, M)$ degrade substantially?

   Test Activation Patterns. Confirm that these same sub-network activations systematically appear for transformed inputs $t(x)$.

VI. **Compare Models.** If Models A and B both excel on raw benchmark scores, but Model A shows a more stable, localized concept sub-network and robust $R(C, M)$ under transformations, it can be argued to exhibit stronger Computational Comprehension of C.

6.5 Illustrative Example

Consider a concept $C_{\text{arithmetic}}$ = *compute the sum of two numbers*. Suppose our domain X is textual inputs like *What is 123 + 456?* We define T as transformations that (a) restate the query in different wording (*Sum of 123 with 456*), (b) embed the numbers in a story, or (c) add noise tokens that should not affect the sum. A model that *really* internalizes $C_{\text{arithmetic}}$ would consistently produce 579 even if the query is rephrased or placed in a distracting context.

Applying mechanistic interpretability, we might discover specific neurons or attention heads that reliably activate for arithmetic queries. Ablating them

should degrade addition performance across transformations. By empirically verifying such a subnetwork, we strengthen the claim that the model holds a stable, concept-like mechanism for addition. Detailed comprehensive illustrations are in Appendix C.

7 Conclusion

Have we truly advanced toward *machine understanding*, or merely built increasingly clever parrots? Can high scores on human-curated tasks ever capture the depth of a concept-driven architecture? And how do we know if a model's seeming mastery reflects stable conceptual knowledge rather than ephemeral statistical tricks? These questions anchor our reimagined path for AI research and practice, informed by a shift from anthropocentric definitions of understanding to more computationally grounded notions.

From Philosophical Inquiry to Applied Methodology

Throughout our exploration, we showed how philosophical debates regarding sensory modalities and subjective experience caution us against overly generous attributions of understanding. At the same time, a strictly anthropocentric frame is unhelpful for analyzing computational systems unlike ourselves. By introducing Computational Comprehension, we direct focus to the operational consistency of concepts represented in neural architectures—a definition testable in practice. Mechanistic interpretability then transforms what once seemed like philosophical abstraction into tangible questions about how models store, manipulate, and apply concepts.

Mechanistic Interpretability as Our Empirical Lens

Mechanistic interpretability serves as the bridge between theoretical claims and observable model behavior, exposing sub-networks or circuits most relevant to a given concept. This approach lets us analyze whether robust conceptual processing is taking place, or whether we are witnessing superficial memorization patterns. When merged with transformation tests, interpretability offers a more holistic assessment of how well a model stabilizes conceptual knowledge across contexts—an empirical window into deeper AI cognition.

Reframing Benchmarks: Strengthening Our Evaluations

The community's current benchmark-focused environment, often driven by an arms race of difficulty , risks conflating performance with legitimate understanding once models surpass human scores. Instead, we propose measuring how well a model preserves conceptual invariants under transformations that should not alter conceptual meaning. This approach valuably shifts emphasis to *how* models achieve their results, rewarding internal coherence over rote data absorption.

Expanded Future Work in Multi-Modality

Our analysis focuses on large language models that primarily process text. However, *multimodal* systems integrating vision, audio, or other input streams introduce new challenges and opportunities for Computational Comprehension. Concepts rooted in visual or sensor-based data (e.g., object permanence, event segmentation) may be tested with transformations that alter viewpoints, backgrounds, or partial occlusions. At the same time, mechanistic interpretability becomes more complex, potentially demanding domain-specific circuit-mapping techniques. Leveraging these modalities could enrich how we define transformations and invariants, but the methodological burden of discovering and ablating relevant circuits also increases.

Taken together, these advances unveil a clearer vista for AI: a future where systems are judged not only by impressive performance metrics but by their demonstrated capacity to encode, transform, and reason with concepts in a sustained manner. By anchoring our evaluations in Computational Comprehension, complemented by mechanistic interpretability and transformation-based testing, we move beyond superficial achievements toward a deeper, more transparent, and ultimately more meaningful form of machine intelligence.

Limitations

Though this paper advocates for a more robust notion of Computational Comprehension and outlines an interpretability-based methodology for uncovering concept-centric representations, a number of limiting factors remain:

Scalability of Interpretability. Modern large language models often feature hundreds of billions

of parameters, making mechanistic analyses at neuron or circuit level extremely challenging. While smaller models or targeted subsets of parameters can be examined, interpreting a full-scale LLM remains computationally intensive and technically difficult, potentially restricting such analyses to well-resourced labs.

Concept Definition Dilemma. Despite the formal approach of modeling a concept as a function or relation, actually specifying certain high-level or context-dependent concepts (e.g., irony, advanced logical reasoning, figurative language) is notoriously difficult. Oversimplified definitions risk overlooking emergent behaviors and interdependencies that cannot be neatly captured by functional mappings alone.

Resource Constraints and Practical Feasibility. Applying repeated transformations, running ablation experiments, and mapping concept circuits can all be computationally expensive. This requirement may slow or hinder widespread adoption of concept-based evaluations, particularly in industry or smaller academic groups where hardware and time are limited. While automated interpretability pipelines are an active area of research, truly large-scale application remains a formidable challenge.

Philosophical Caveats. Finally, even the most thorough demonstration of stable concept circuits does not prove the presence of conscious or subjective states. Our notion of Computational Comprehension is deliberately *non*-anthropocentric; it makes no claims about qualia or mental experiences. This helps avoid overattribution of human-like cognition to machines, yet it may leave some philosophical critiques unresolved, especially those that posit experience as central to genuine understanding (e.g., Mary's Room).

These limitations do not diminish the value of Computational Comprehension as a powerful conceptual and evaluative tool; rather, they highlight the need for further research, methodological refinement, and interdisciplinary collaboration in continuing to push the boundaries of *machine understanding*.

Ethics Statement

This work proposes a framework for examining and measuring Computational Comprehension in large language models using mechanistic interpretability and transformation-based evaluations. While our aim is to foster deeper insights into a model's conceptual capacities, we acknowledge several ethical considerations:

- **Misuse of Interpretability.** Enhanced methods for dissecting internal model mechanisms, though beneficial for transparency, may be misapplied to extract proprietary information or target sensitive aspects of model behavior. Researchers and developers should exercise care in deciding which details are publicly disclosed, balancing transparency with the potential for malicious exploitation.

- **Bias and Fairness.** As we refine tools to identify conceptual circuits, it is equally important to ensure they detect and mitigate biases. Models can unintentionally encapsulate harmful stereotypes that escape surface-level benchmarking. Mechanistic interpretability thus has an ethical imperative to uncover and address such issues more systematically.

- **Data Privacy.** Many interpretability protocols rely on intensive probing or ablation experiments with curated datasets. Researchers should remain cognizant of privacy considerations and legal constraints—particularly when dealing with personal or sensitive data—and follow best practices to anonymize or redact identifying information.

- **Anthropomorphizing and Responsibility.** While Computational Comprehension provides a non-anthropocentric lens, there remains a risk of over-attributing agency or accountability to models. We clarify that these evaluations do not confer moral or legal responsibility on the system and should not be used to justify autonomy or moral standing for LLMs.

- **Potential for Unequal Access.** Robust interpretability frameworks can be computationally and logistically expensive, raising concerns that only a small segment of the AI community will have sufficient resources to conduct these studies. The field should encourage shared resources, open-source tools, and collaboration to ensure that concept-based evaluations do not reinforce inequitable access or hinder responsible innovation.

By presenting Computational Comprehension as an additional source of insight, our goal is to

complement, rather than displace, existing ethical and regulatory frameworks. We advocate for open, interdisciplinary dialogue on how best to leverage detailed model understanding to enhance both performance and societal benefit, mindful of the risks of misuse, bias, or inequity.

References

Daman Arora, Himanshu Gaurav Singh, et al. 2023. Have llms advanced enough? a challenging problem solving benchmark for large language models. *arXiv preprint arXiv:2305.15074*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *FAccT*.

Michael Timothy Bennett. 2023. On the computation of meaning, language models and incomprehensible horrors. In *International Conference on Artificial General Intelligence*, pages 32–41. Springer.

Carl Bereiter. 2002. *Education and Mind in the Knowledge Age*. Routledge.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety - a review. *arXiv*.

Nigel Collier, Fangyu Liu, and Ehsan Shareghi. 2022. On reality and the limits of language data: Aligning llms with human norms. In *Annual Meeting of the Cognitive Science Society*.

Lucia Foglia and Robert Wilson. 2013. Embodied cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

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

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2023. Think before you speak: Training language models with pause tokens. *arXiv preprint arXiv:2310.02226*.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jos´e Hern´andez-Orallo. 2020. Ai evaluation: On broken yardsticks and measurement scales.

Frank Jackson. 1986. What mary didn't know. *The Journal of Philosophy*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Raphaël Millière and Charles Rathkopf. 2024. Anthropocentric bias and the possibility of artificial cognition. In *ICML 2024 Workshop on LLMs and Cognition*.

Philipp Mondorf and Barbara Plank. 2024. Beyond accuracy: Evaluating the reasoning behavior of large language models - a survey. *ArXiv*, abs/2404.01869.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jesse Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *ICLR*.

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

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

P Rajpurkar. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

John R Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

James Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8:1–11.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Alex Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Zeping Yu and Sophia Ananiadou. 2024a. Interpreting arithmetic mechanism in large language models through comparative neuron analysis. *arXiv preprint arXiv:2409.14144*.

Zeping Yu and Sophia Ananiadou. 2024b. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3267–3280.

Pardis Sadat Zahraei and Ehsaneddin Asgari. 2024. Turingq: Benchmarking ai comprehension in theory of computation. *EMNLP Findings*.

Wei Zhang, Chaoqun Wan, Yonggang Zhang, Yiu-ming Cheung, Xinmei Tian, Xu Shen, and Jieping Ye. 2024. Interpreting and improving large language models in arithmetic calculation. *arXiv preprint arXiv:2409.01659*.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

Appendix

A Preliminary Exploration of Transformations for Logic

For logical inference tasks, we define a relation $C(\phi, \psi)$ as a function returning `true` if ψ is logically entailed by ϕ in propositional logic. To test transformational robustness, one can apply rewritings or reorder terms without altering the underlying logical dependencies. Examples include variable renaming, distribution of negations, or rewriting implications (e.g., $p \rightarrow q$ as $\neg p \vee q$). If M displays high consistency under such transformations, it suggests the model holds a more generalizable representation of logical entailment, instead of merely memorizing surface patterns.

B Further Discussion of Mechanistic Ablation Techniques

Main text discussions of ablation (Section 3) can be expanded by referencing specialized protocols such as *Neuron-Level Knowledge Attribution in Large Language Models*. These involve:

1. Identifying neuron groups strongly correlated with a target concept based on their activation patterns.

2. Zeroing out or injecting noise into those neuron activations while the remaining network remains intact.

3. Measuring the consequent drop in performance on tasks or test suites that rely on the targeted concept.

Marked degradation suggests that knowledge of the concept is at least partially localized in those neurons. However, the possibility of distributed or redundant representations must be carefully weighed, potentially requiring multiple ablation variants or iterative neuron-discovery steps.

C Illustrative Example: Arithmetic Comprehension

Consider a concept $C_{\text{arithmetic}}$ = *compute the sum of two numbers*. This concept provides an ideal test case for Computational Comprehension due to its well-defined structure and clear expected outputs.

Concept Definition and Input Domain

Our domain X consists of textual inputs requesting arithmetic operations, such as *"What is 123 + 456?"*. The concept function $C_{\text{arithmetic}}$ maps these inputs to their numerical solutions (in this case, 579).

Transformation Family

We define transformation family T to include operations that preserve the underlying arithmetic concept while varying surface presentation:

1. **Paraphrasing:** Restate the query using different wording (*"Calculate the sum of 123 and 456"*, *"What do you get when you add 123 to 456?"*)

2. **Contextual Embedding:** Place the arithmetic operation within a narrative context (*"John had 123 apples and received 456 more. How many apples does John have now?"*)

3. **Noise Addition:** Insert irrelevant tokens that should not affect the computation (*"What is, um, let me think, 123 plus, you know, 456?"*)

4. **Format Variation:** Present numbers in different formats (*"What is one hundred twenty-three plus four hundred fifty-six?"*)

Evaluating Computational Comprehension

A model that truly comprehends the concept of addition would consistently produce the correct sum (579) across all these transformations. This consistency indicates that the model has internalized the abstract operation of addition rather than merely memorizing specific input-output patterns.

Mechanistic Analysis Protocol

Recent work in mechanistic interpretability has demonstrated methods for identifying arithmetic circuits in language models. We propose the following protocol to verify computational comprehension of arithmetic:

1. **Circuit Identification:** Using causal mediation analysis or activation patching, identify the specific neurons, attention heads, or MLP modules that activate when processing arithmetic queries. Research suggests these components often reside in mid-sequence early layers and the final token's later layers.

2. **Activation Pattern Analysis:** Record activation patterns across the transformation family T. A model with true computational comprehension should show consistent activation in the same sub-networks regardless of how the arithmetic query is presented.

3. **Targeted Ablation:** Temporarily disable the identified arithmetic circuit components through zero ablation or mean ablation. If these components are truly responsible for arithmetic processing, performance should degrade specifically on arithmetic tasks while leaving other capabilities intact.

4. **Cross-Task Comparison:** Compare activation patterns during arithmetic tasks with patterns during number retrieval tasks and factual knowledge questions. True arithmetic comprehension should exhibit distinct circuit activation compared to mere number recognition or general knowledge retrieval.

Expected Observations

If a model possesses computational comprehension of arithmetic, we would expect to observe:

1. Consistent performance ($M(t(x)) \approx C(x) = 579$) across all transformations $t \in T$

2. Identifiable sub-networks that activate specifically for arithmetic operations

3. Significant performance degradation on arithmetic tasks (but not other tasks) when these sub-networks are ablated

4. Similar activation patterns across different presentations of the same arithmetic problem

This arithmetic example demonstrates how Computational Comprehension goes beyond surface-level accuracy to examine the internal mechanisms supporting concept representation and manipulation. By verifying both transformational robustness and the presence of dedicated conceptual circuits, we can distinguish models that truly comprehend arithmetic from those that merely pattern-match on specific input formats.