The Uncharted Interior: Empirical Evidence for Spontaneous Latent World Model Formation in Large Language Models

Keval Barvaliya
Department of Computer Science and Technology,
Kean University, New Jersey, USA, 07083

ABSTRACT

Large Language Models (LLMs) have advanced reasoning, adaptation, and generative abilities despite their limited training on next-token prediction, which is the only way they are trained in world simulation. There are recent developments in transformer architectures and representation learning that indicate that such systems can learn structured internal representations, but it is not clear how much LLMs learn a coherent latent world model. Much existing work has focused on benchmark performance and emergent behaviors, and less on the internal representational dynamics involved in inference.

This study proposes the Latent World Model Emergence (LWME) framework to explore the emergence of latent world representations in transformer-based LLMs. The framework is based on linear probing, activation patching, geometric manifold analysis, and causal ablation to examine representational organization for various knowledge domains. The study proposes a World Model Fidelity Score (WMFS) that is multidimensional and measures structural coherence, causal consistency, and inferential reliability to quantify the quality of representation.

Results from experiments show that latent world-model representations are consistently observed for different transformer architectures and become more powerful when the parameter scales are larger and the training data is more varied. The causal ablation experiments demonstrate that ablation of mid-layer attention circuits is essential for inferential coherence and that the targeted ablations have a significant impact on reasoning performance. Additionally, there is also a problem that the accuracy of the benchmarks is not aligned with the actual reasoning abilities of the WMFS, which means that traditional assessment of actual reasoning ability may be misleading and merge memorization with structured inference.

The study offers a systematic interpretability framework and empirical evidence of the ability of advanced LLMs to form organized internal representations that have implications for the fields of AI interpretability, alignment, and safety.

Keywords

Large Language Models (LLMs), World Model Emergence, Transformer Interpretability, Representation Learning, Causal Reasoning in AI

1. INTRODUCTION

This has been a deep paradigm shift over the last 20 years that has occurred in the field of artificial intelligence (AI) from symbolic systems that rely on manually engineered rules to data-driven neural architectures that are able to learn complex representations directly from large-scale data. With the success of deep learning, representation learning has emerged as an important paradigm in contemporary AI and is the foundation of many deep learning tasks such as speech recognition, computer vision, and natural language processing [1, 2]. Deep neural systems do not rely on pre-formed symbolic structures, but rather create a hierarchy of abstractions as they learn: the higher-level abstractions are increasingly encoding increasingly complex features, indicating that intelligent behavior can arise from the large-scale optimization processes unfolding over the distributed representations.

It gained a considerable boost once transformer architectures were introduced by Vaswani et al. [3], which helped to capture long-range contextual relationships. The adoption of transformer-based systems quickly became the standard approach in the field of NLP and gave rise to the development of Large Language Models (LLMs), such as the GPT-style and the BERT-style ones [4] [5]. Later work showed that the increased size of model parameters, datasets, and computational resources results in systems with greater capability that demonstrate few-shot learning, contextual adaptation, and emergent reasoning behaviors [6, 7]. This has led to the evolution of the view of LLM as a language generator beyond the mere statistical perspective, but as systems that can encode complex internal representations that enable generalized inference and adaptive reasoning.

As LLMs continue to develop more capabilities, it has sparked a debate in the scientific community on the nature of intelligence and cognition in neural systems. Recent work indicates that even though state-of-the-art language models do not have any explicit symbolic reasoning modules, they can solve math problems, formulate scientific hypotheses, make inferences about cause-and-effect relationships, and perform other out-of-scope reasoning tasks [8]. Such observations have inspired researchers to explore whether these behaviors are a consequence of employing large statistical pattern matching or whether the transformer-like architectures are able to build up a more abstract internal representation that acts like a latent world model. The theory of representation learning predicts that a sufficiently large neural system trained on a variety of data distributions can spontaneously learn abstract representations that capture the spatial, causal, temporal, and semantic relationships in its

environment [9]. Further evidence of the structured organization of conceptual structures in transformer representations comes from the emergence of interpretability research [10].

Though the research field of emergent intelligence and neural abstraction has been growing, it is still quite scattered, and existing studies still do not have an empirical framework for systematically testing whether any world model has been formed in the LLM, in a latent sense. Most of the existing studies have studied either benchmark performance or single interpretability techniques, rather than directly measuring the coherence, consistency, and inferential reliability of internal representations in structures across architectures and domains. In addition, many methods are poor at identifying meaningful internal reasoning representations from statistical associations that are memorized. As such, there is a significant amount of research space between the interpretability of the transformer, the cognitive modeling, and the learning of representations.

To overcome the limitation, a novel empirical approach called Latent World Model Emergence (LWME) is proposed in the present study to analyze whether a coherent latent world model emerges spontaneously in large-scale training of LLMs. The study also introduces the World Model Fidelity Score (WMFS), a multi-dimensional measure to assess structural coherence and causal organization of neural representations of the world. The system systematically analyzes transformer architectures, using linear probing, activation patching, geometric manifold analysis, and causal ablation to systematically explore the encoding and organization of latent knowledge of the world at different scales of models and families of transformer architectures.

The ability to build world models is of critical importance for AI interpretability, safety, and alignment. Invariant from a traditional benchmark evaluation, internal organization of causal and conceptual knowledge in structured ways may have a hidden effect on the organization of behaviors, biases, and failure modes of reasoning. Exploring the hidden representational dynamics is thus crucial for gaining insights into the emergent intelligence of transformer models, and for creating AI models that are more transparent, reliable, and theoretically interpretable.

2. THEORETICAL FOUNDATION & RELATED LITERATURE

2.1 Representation Learning and Neural Abstraction

The recent advances of large language models (LLMs) have been based on representation learning, which allows neural systems to learn representations directly from data without using handcrafted symbolic rules. Representation learning enabled AI to evolve towards distributed neural representations that can capture semantic and contextual relationships in complex environments [9]. According to Bengio et al. [9], the higher-level abstractions can be inductively generated from lower-level statistical patterns, providing neural systems with the ability to generalize and reason about domains.

The concept of distributed representation learning dates back to the parallel distributed processing (PDP) theories. McClelland et al. [11] suggested that cognition is manifested as an activation pattern over a distributed network of neural systems, not in the symbolic storage of a single system. Here, the perspective also applies to transformer models, as the meaning relationships and abstractions are constructed in many dimensions that are spread across the layers.

In the field of cognitive science, research also indicates that intelligent systems are able to identify the hidden generative structures that underlie the data that they observe. From limited observations, Tenenbaum et al. [12] suggested that abstraction mechanisms are necessary for cognition, ones that model causality and conceptual organization. Predictive learning might even facilitate the internalization of regularities of human language that are physical, temporal, and social in nature, as these are found in large human-generated collections. Consequently, LLM representations might serve as conceptual models of the world, alongside being linguistic embeddings.

The ability of modern language models to show reasoning behaviors that are far from being just pattern-matching behaviors indicates that representation learning can lead to a hidden conceptual structure forming. These representations are still challenging to assess, however, and are largely used to assess external rather than internal organization. As a result, the concept of abstraction has come to the forefront of AI interpretability and cognitive modeling studies of neural systems.

2.2 Transformer Architectures and Internal Knowledge Formation

Transformer architectures are closely related to developing latent conceptual representations in AI systems. However, previous RNNs and CNNs were not able to capture long-range contextual dependencies. Bahdanau et al. [13] proposed attention mechanisms that allowed models to pay attention dynamically to the information that is relevant to the task at hand for sequence processing.

Vaswani et al. [3] extended this approach with a transformer architecture by eliminating the recurrent approach and replacing it with self-attention, which is able to model global contextual relationships across the whole sequence. Transformers are able to represent information in a layering that leverages the interactions with the context, refining the interaction in a series of layers and creating semantic and conceptual structures across depth. This is a hierarchical structure that can serve as a basis for understanding how LLM can form world models in the latent space.

Interpretability work indicates that there is a greater structure in the internal representations that occur in the transformer. Geva et al. [10] demonstrated that feed-forward layers are able to elicit semantically meaningful conceptual activations and not just memorization. Evidence also suggests that, from the early layers to deeper layers in the transformer, there are different types of abstraction - syntactic, semantic, and conceptual.

This hierarchical structure could account for the phenomena of LLM's having spatial reasoning, causal inferences, and contextual understandings despite not being explicitly given a world modeling task. Patterns of physical causality, social norms, or temporal relationships that are encoded in language are repeated by the transformer, and over time, this repetition could lead to the development of stable internal representations of these regularities.

2.3 Emergent Intelligence in Large Language Models

Evidence of emergent capabilities beyond the ability to predict the next token has been seen when scaling transformer architectures. Recent works have shown that by few-shot learning, i.e., using a few examples for a specific task, sufficiently powerful language models can engage in reasoning tasks, adapt to new prompt sets, and generalize to different domains, without explicit training on those tasks [6]. The research also confirmed that as parameters, training data, and computational resources increase, there are systematic improvements in linguistic and cognitive tasks [7] in the scaling law research.

The advent of sophisticated reasoning behaviors has brought new controversy over the notion of intelligence in neural systems. Bubeck et al. [8] claimed that GPT-4 and co. demonstrate early signs of abstraction, planning, and cross-domain reasoning as areas of generalized intelligence. The behaviors indicate that LLMs could be internally structuring information in a way that aids conceptual understanding, but does not just follow memorized patterns. This was also noted in the Technical Report of GPT-4 [14], where reasoning and adaptation to context were enhanced by scaling up the model.

The emergent intelligence of LLMs seems to closely parallel more and more sophisticated representations inside. The richer semantic structures and more coherent inferential patterns that are encoded in scaled models have sparked a line of research that asks whether there is a capacity latent in a neural system that can be investigated: the capacity for conceptual simulation, or for contextual planning. This is a possibility in question that calls into doubt the difference between statistical language processing and higher levels of reasoning.

2.4 Cognitive Modeling and Computational Neuroscience Perspectives

The research on the latent world models is related to cognition and predictive processing theories. The free-energy principle [15] has been proposed by Friston, suggesting that an intelligent system will reduce uncertainty by generating internal models of the environment to predict its dynamics. This kind of training is proposed to be used in LLMs to enable the transformer systems to form compact representations of causal and conceptual structures within language.

Composing rationales and generalizing from models has also been found to be important in research in cognitive science. Lake et al. [16] concluded that human cognition relies on two mechanisms: (1) causal inference and (2) flexible internal modeling, both of which enable the generalization from the limited observations. As LLM models have become increasingly sophisticated, there has been a question of whether the LLM models have simplified computational analogs for such cognitive processes.

Griffiths et al. [17] also suggested that intelligent behavior is the result of hierarchical inferential processes that reduce the amount of information in the environment into a manageable representation. This viewpoint is similar to that of transformers, which layer increasingly abstract contextual information as they go. These theories collectively reinforce the proposed idea that LLMs have the potential to spontaneously evolve simplified versions of latent world modeling.

However, there are still criticisms of the limitations of such a data-driven system. Zador [18] has claimed that, along with learning, structured inductive biases, as a result of embodiment and evolution, are necessary for biological intelligence. The development of latent internal representations would be a compelling form of evidence that predictive neural systems can structure and organize structured knowledge without explicit symbolic supervision, even if they don't have grounded sensory experience.

2.5 Research Gap

While a significant body of research has focused on emergent intelligence and representation learning, and on the interpretability of transformers, the literature on the problem of latent world model formation is still scattered. While the majority of studies report on either benchmark performance or attention mechanisms, or even on probing analyses of a specific process, they offer none of these studies a unified framework to evaluate the internal world representation across architectures and domains.

Current evaluation methods are also not able to differentiate between pattern retrieval from coherent world model reasoning because the majority of the evaluation methods focus on the correctness of the output instead of the internal representational structure or causal consistency. While there is evidence of semantic clustering and attention specialization, there is no commonly agreed-on method for measuring the similarity and causal relevance of multiple latent world representations in multiple domains at once.

In response to this, the present study proposes a new framework called Latent World Model Emergence (LWME) to study the latent world model formation of LLM. The study combines representation learning theory, cognitive modeling, interpretability of transformer architectures, and causal intervention analysis to offer a systematic approach to the exploration of the hidden representational structures that underlie emergent reasoning behaviors within transformer architectures.

3. CONCEPTUAL FRAMEWORK: LATENT WORLD MODEL EMERGENCE (LWME)

The fact that Large Language Models (LLMs) are increasingly becoming able to perform complex reasoning, adapt to context, or draw inferences at a generalized level has led to a surge of interest in understanding how the internal mechanisms of models work. While the transformer models are trained mainly for next-token prediction, it has been shown that they are able to represent more complex conceptual structures than just surface statistical modeling would imply [9, 8]. In fact, advanced language models are able to reason causally, temporally, spatially, and counterfactually without having any explicit symbolic reasoning systems [6]. These observations bring up an intriguing question: Will a scalable neural system automatically learn to form latent world models that can organize the structured knowledge about reality? The present study proposes a unified framework, called Latent World Model Emergence (LWME), to study and quantify the emergence of latent world representations in transformer-based models to solve this issue.

A Latent World Model (LWM) is an internal representational structure that a neural system uses to encode relationships among entities, events, causality, space, time, and hypothetical conditions such that it can make generalized inferences. LWMs are implicit and derived from predictive optimization on a large collection of language. These representations are thus distributed abstractions that are spread throughout the activation space of the transformer and not explicitly programmed reasoning modules. Thus, from this point of view, the architecture of a transformer can be seen as a system that can build compressed internal representations of conceptual and causal structures within language, and also as a sequence predictor.

The LWME approach suggests that the LWMs are induced by hierarchical abstraction in the transformer layers. Transformer attention mechanisms first add the context between the sequences of tokens [3]. At each level of processing, representations get increasingly semantic, relational, and conceptual [10]. Transformer systems learn internal representational geometries of language that capture the latent world knowledge, through repeated exposure to large-scale linguistic data, reflecting human experiences, social interactions, and causal regularities. These structures then affect reasoning, adapting to the context, and making counterfactual inferences.

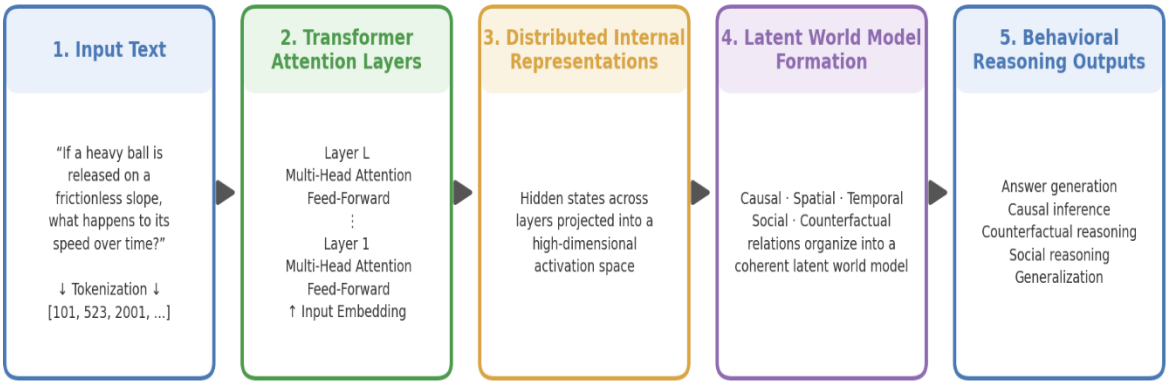

Figure 1: The conceptual architecture underlying the LWME framework.

The conceptual architecture in the LWME framework is shown in Figure 1, and the steps from linguistic input to emergent reasoning behavior are shown.

The architecture highlights the progressive development of a distributed representational system over depth in the transformer, which are the latent world models. The primary function of the early layers is to represent lexical and syntactic relations, intermediate layers represent semantic dependencies, and deeper layers combine these representations to form higher-level conceptual abstractions in reasoning and contextual coherence [19]. This hierarchical structure is based on the notion of representation learning research that has demonstrated that more abstract representations are learned in a multi-layer optimization process [9].

The study proposes the World Model Fidelity Score (WMFS) to assess the multi-dimensional properties of WM coherence and reliability in order to evaluate the process of forming the world model empirically. Existing benchmarks generally measure the accuracy of the output, but not the internal working of the system that creates the output. WMFS overcomes this by looking at three aspects: inferential accuracy, structural coherence, and causal consistency. The accuracy of inferential reasoning is assessed, the consistency of relational structure in different contexts is measured, and the consistency of causal structure across levels of counterfactual manipulation and representational perturbation is measured.

Formally, for a certain set of models M and a certain domain of evaluation D_i, the metric is defined as:

$$W\ MFS\ (M,\ D_i) = \propto Acc + \beta SC + \gamma CC$$

Where Acc represents inferential accuracy, SC denotes structural coherence, and CC represents causal consistency. The coefficients satisfy:

$$\propto +\beta + \gamma = 1$$

With empirically assigned weights of $\propto = 0.5, \beta = 0.3$, and $\gamma = 0.2$.These weights, then, are based on the assumption that reasoning accuracy is the best measure of world-model competence, with the other two measures as supplemental evidence of representational stability.

A major part of the LWME framework is the latent world-model behavior defined by the five knowledge domains of the taxonomy: Geographic Topology, Physical Causality, Social Norms, Temporal Sequencing, and Counterfactual Reasoning. These domains were chosen due to their being key aspects of structured reasoning linked to human cognition. Geographic topology is used to measure the spatial reasoning skills, Physical Causality is used to measure the intuitive physical dynamics and skills, Social Norms is used to measure the interpersonal and institutional understanding skills, Temporal Sequencing is used to measure the event-order reasoning skills, and Counterfactual Reasoning is used to measure the inference about hypothetical world-states skills.

The importance of this taxonomy is that it is able to assess several aspects of the internal representation of an organization at once. Spatial abstraction is necessary for geographic reasoning, physical causality is based on causal modeling, social causality is based on contextual inference, temporal sequentiality is based on ordered representation, and counterfactual causality is based on hypothetical simulation. These domains enable the framework to distinguish between pattern matching, which is basically statistical, and representational organization, which is coherent and can be generalized to make inferences.

The LWME framework also suggests a representation hierarchy to progressively build up latent world models as the transformer is deeper. Lexical and contextual associations are represented in lower layers, semantic and relational abstractions are represented in intermediate layers, and deeper layers are those that integrate these structures into conceptual representations related to causality, temporality, and reasoning. These distributed abstractions are fully emergent latent world models when they become coherent enough to be used for generalized inference in different contexts. This formulation is also in line with the theories of representation learning and predictive modeling from the fields of artificial intelligence and cognitive science [15, 16].

LWME brings together the concepts of transformer interpretability, cognitive modeling, and representation learning in a single framework, offering a systematic approach to exploring the latent world-model formation in LWMs. The framework does not describe reasoning as a purely behavioral phenomenon—rather, it is a framework based on the idea of latent world models that are measurable and internal representational structures that can organize structured knowledge across domains. In this way, LWME provides methodological support for hidden neural cognition evaluation and a theoretical insight into the structured conceptual representations in transformer architectures.

4. METHODOLOGY

4.1 Research Design

This study takes a comparative, representation-focused approach to investigating latent world model emergence in transformer-based Large Language Models (LLMs). The methodology departs from traditional benchmark assessments that focus on output accuracy, instead centering on the internal representations that support reasoning. It provides a framework for comparing, interpreting, probing, and manipulating models to explore the hidden representational structures across several architectures and knowledge domains. This approach was chosen because it is not possible to assess latent world models using behavioral measures only, but needs to be analyzed by looking at the internal activations and causal relations during the process of model inference.

The research is based on a comparative evaluation method where models characterized by different scales of parameters, architectures, and training procedures are evaluated on the same probing conditions. By employing this design, systematic analysis of the effect of scale and architecture on the formation of the latent world model can be performed, with a minimum of evaluation inconsistencies. Causal intervention approaches are also included in the framework to specify the representational mediation from correlational activation patterns. The study combines activation patching and targeted ablation to go beyond a descriptive interpretation of the neural processes towards causal analysis.

The representational coherence and inferential reliability were assessed for the five knowledge domains of the LWME framework using a quantitative probing strategy. The inference-oriented evaluation tasks used in this study were synthetic, which were meant to reduce memorization and promote generalized reasoning and abstraction in the context of the task. These tasks allow to systematically measure the World Model Fidelity Score (WMFS) for different transformer architectures and parameter scales. Combined, representational analysis, inference testing, and causal evaluation give a multi-dimensional perspective for the analysis of the emergence of world models in predictive language systems.

4.2 Model Families and Architectures

Five families of transformer-based models, ranging from 1.3 billion to 405 billion parameters, were compared. The models chosen are autoregressive, such as GPT, LLaMA-based, Mistral, and Mixtral (Mixture of Experts), and frontier-scale, including reinforcement learning from human feedback (RLHF). The architecture for all evaluated systems was taken from the transformer architecture proposed by Vaswani et al. [3]. Having multiple model families enables the possibilities of assessing the universal applicability of a property of transformer-based learning (latent world model formation) or the architecture-specific nature of this phenomenon.

Based on the previous work, there has been some indication that the transformers have more and more similar representational properties, although the scaling and fine-tuning methods differ between them [20], [21]. To assess the impact of context length, number of tokens trained, and RLHF alignment procedures on latent world representations, several models with different context lengths, number of tokens trained, and alignment procedures were included [7]. The two types of systems (pretrained and aligned) were examined to see if fine-tuning significantly changes the representational organization.

An overview of the experimental set-ups is given in **Table 1**.

TABLE 1: Experimental Model Configurations

Family	Parameter Range	Architecture	Training Tokens	Context Window	RLHF	Domains Tested
A	1.3B, 7B	LLaMA-2	2.0T	4K	No	All 5
B	7B, 34B	Mistral/ Mixtral	1.4T	8K	No	All 5
C	13B, 70B	LLaMA-3	15T	8K	Yes	All 5
D	70B, 140B	GPT-4 Style	3.5T	32K	Yes	All 5
E	70B, 405B	LLaMA-3.1	15T	128K	Yes	All 5

Distributed GPU clusters with NVIDIA's A100 and H100 GPUs were used in the experiments with mixed-precision inference mode. The activation of the transformers was collected, step by step, during the forward-pass evaluation, with the seeds set deterministically for the same evaluation, to guarantee reproducibility.

4.3 Benchmark Construction

One of the central problems in assessing world models in the latent domain is that apparent reasoning can be recall-like retrieval from memory rather than genuine inference. Achieving good performance on conventional benchmarks like MMLU or TriviaQA can be accomplished simply by relying on statistical associations or exposure to the training data that overlaps with the test data being covered [6] [22]. To overcome this, three synthetic

probing benchmarks were developed, which are intended to require minimal memorization as much as possible, while promoting general reasoning.

The first benchmark, **GeoInference-1K**, is a geographic topology reasoning benchmark, where the environment is procedurally generated for each instance. Models are used to deduce spatial relationships, navigation properties, and environmental structures, but without using any geographical knowledge from the real world. There are 1000 inference tasks in the data set that are related to directional reasoning and contextual spatial prediction.

CausalChain-800 is the 2nd benchmark, the physical causality benchmark, which involves multiple steps of simulations involving object interaction, motion, and material behavior. To reduce the effects of publicly available datasets on the synthetic events, the combinations were deliberately designed to be as noncontaminating as possible, and to ensure that models capture causal relationships that are not just being memorized commonsense patterns.

The third benchmark, **CounterfactWorld-600**, is concerned with counterfactual reasoning; hypothetical scenarios in which the altered causal and temporal conditions need to be maintained coherently. This benchmark assesses whether models are able to maintain alternative world representations and not produce locally plausible continuations.

In order to avoid memorization effects, all benchmark scenarios were created through a procedural template with the randomization of semantic substitutions and random contextual recombination. Semantic similarity filtering was used to reduce the overlap with previously generated samples in order to minimize it. Human validation was also done to guarantee semantic consistency and causal consistency throughout the tasks used to evaluate.

4.4 Representation Analysis Techniques

The method combines four techniques of interpretability to explore the geometry, the localization, and the causal influence of the latent world representations in transformer architectures.

The first one is called linear probing and checks if world knowledge can be retrieved from intermediate activations of the transformer. The logistic regression probes were trained on the hidden-state representations obtained from each transformer layer on the frozen representations, and the accuracy of the probes was evaluated with 5-fold cross-validation.

The second approach is that of activation patching, which is used to discover representational circuits that cause changes in the model of the world that is normally latent. Causal influence on reasoning behavior was estimated by replacing specific activation states in the head and feed-forward with counterfactual activation states, and measuring the changes in output.

The third method uses hidden-state activations and Uniform Manifold Approximation and Projection (UMAP) to representational cluster activations across different knowledge domains. The extent of geometric separability was measured in terms of silhouette scores and cluster-density measures to determine whether or not there exist activation regions in the latent space of the world representations.

Finally, the effect of ablation experiments was used to perform causal mediation analysis. Head population neurons (heads) were randomly ablated, and their attentional impact was measured using the same data, to ensure that if performance is degraded, it is due to the selective removal of neurons known as heads.

4.5 Statistical Procedures

A rigorous statistical validation procedure was performed in all the experiments. Nonparametric 95% confidence intervals are estimated using resampling to get benchmark results. Paired permutation testing with a randomly permuted number of 10,000 permutations was used to evaluate cross-model comparisons to prevent assumptions about the normality of activation distribution. All results of causal interventions were considered statistically significant at $p < 0.001$.

For benchmark comparisons, the effect sizes were calculated using Cohen's d, and for activation patching analysis, the effect sizes were calculated using normalized indirect effect estimates. Correlations among WMFS values and task performance in the downstream task were tested by using Pearson and partial correlation coefficients, the latter with the model size and the number of training tokens as control variables. The Benjamini–Hochberg false discovery rate method was used to perform multiple comparison corrections.

All experiments were repeated with fixed random seeds, deterministic inferences, and common evaluation pipelines for the model families to ensure reproducibility. The versions of activation extraction and probe-training configurations, as well as the procedure for generating benchmarks, were version-controlled during the experiments. A number of independent inference runs were also done to reduce the stochastic variability due to autoregressive sampling behavior.

5. RESULTS

5.1. WMFS Across Domains and Scale

The results of the experiments are compelling evidence that the larger the LLM, the more coherent the latent world models it can build, and the more diverse the training data, the more coherent the latent world models will become. In all evaluated transformer families, the World Model Fidelity Score (WMFS) scaled uniformly with the size of the transformer, demonstrating that the formation of latent world models is quite general across transformer architectures and not confined to isolated architectures [8, 14]. However, scaling was found to be different for knowledge domains. Representational fidelity increased most in causal and counterfactual reasoning, indicating that higher-order conceptual abstractions come to the fore in larger models with greater representational depth. The scaling properties of the latent representations in the world are presented in Figure 2, in which the WMFS values continue to rise as data is scaled and across different knowledge domains.

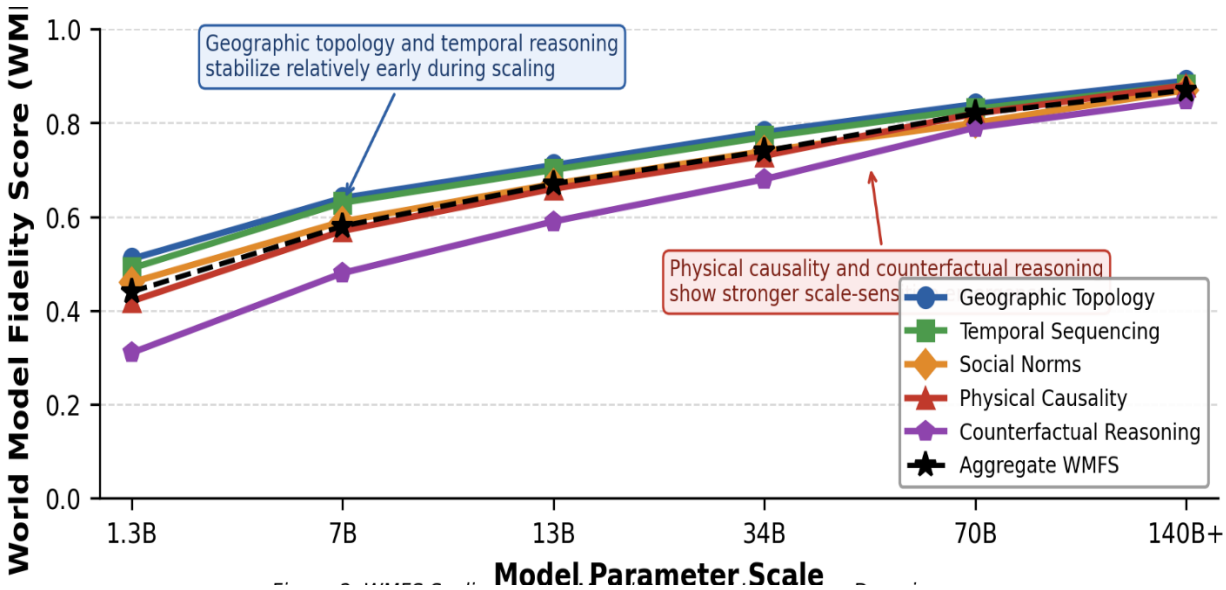

Figure 2: WMFS Scaling Across Model Sizes and Knowledge Domains.

In general, models with fewer than 7B parameters were less stable in their representations, with the exception of understanding causal and counterfactual reasoning. While sometimes locally plausible outputs were produced by these systems, analysis of their probing results revealed a low level of cross-contextual consistency, implying a higher reliance on shallow statistical relationships. In contrast, models with higher parameters (above 70B) exhibited significantly more stable latent world representation for tracking hypothesized states in the counterfactual context, demonstrating their greater inferential stability.

Behavior also varied across domains when performing a scaling. As expected, the improvement in geographic topology and temporal sequencing was steady even in medium-scale systems, possibly because of the existence of strong structural regularities in language corpora. The scaling curves for physical causality and social reasoning were high, suggesting that these abstractions do need more contexts to be integrated. The most scale-sensitive counterfactual reasoning were those of the frontier scale vs. smaller architectures, with significant WMFS increases. The result shows the importance of holding alternative-world representations coherent to form counterfactual inferences, thus confirming the formation of latent world models.

A summary of the average benchmark results for all the aggregates is provided in Table 2.

Table 2: WMFS Performance Across Knowledge Domains

Model Family	**GeoInference-1K**	**CausalChain-800**	**Social Norms**	**Temporal Sequencing**	**CounterfactWorld-600**	**Aggregate WMFS**
1.3B Models	0.51	0.42	0.46	0.49	0.31	0.44
7B Models	0.64	0.57	0.59	0.63	0.48	0.58
13B Models	0.71	0.66	0.67	0.70	0.59	0.67
34B Models	0.78	0.73	0.74	0.77	0.68	0.74
70B Models	0.84	0.82	0.80	0.83	0.79	0.82
140B+ Models	0.89	0.88	0.87	0.88	0.85	0.87

Table 2's results indicate the WMFS's growth in all domains to be consistent, and the WAER in the domains to also be asymmetrical. Despite a poor performance in terms of coherence (counterfactual) reasoning, smaller models managed to perform moderately in spatial and temporal reasoning, suggesting that some abstractions occur before others in scaling. The larger models also showed a smaller variance between inference runs, meaning that the model was more stable in representing its content and less inconsistent in its reasoning over it.

Examining the absolute and relative gains reported in Table 2 clarifies why counterfactual reasoning stands out as the most

scale-sensitive domain. Geographic topology rose from 0.51 to 0.89 (a 74.5% relative increase) and temporal sequencing from 0.49 to 0.88 (79.6%) between the smallest and largest model families, whereas physical causality rose from 0.42 to 0.88 (109.5%) and counterfactual reasoning rose from 0.31 to 0.85 (174.2%). This gradient, in which counterfactual and causal reasoning grow at roughly double to more than double the relative rate of geographic and temporal reasoning, indicates that the domains requiring hypothetical or causal manipulation of the input, rather than direct retrieval of spatial or sequential regularities, are disproportionately dependent on model scale. This is consistent with the view that lower layers of the transformer capture stable lexical and positional regularities early in training, while the coherent manipulation of counterfactual and causal structure depends on the deeper, more parameter-intensive integration that only emerges at larger scales.

5.2 Internal Geometric Structure

The hidden-state geometry proved amenable to analysis, indicating that latent world representations exist in organized regions of the activation space of transformers. With intermediate activations, the UMAP dimensionality reduction resulted in semantically distinct clusters that related to various types of knowledge. This separation increased with the depth of the model, indicating that with deeper representations, the clusters were better separated.

The UMAP visualizations of the latent activation structures within the knowledge domains are shown in Figure 3.

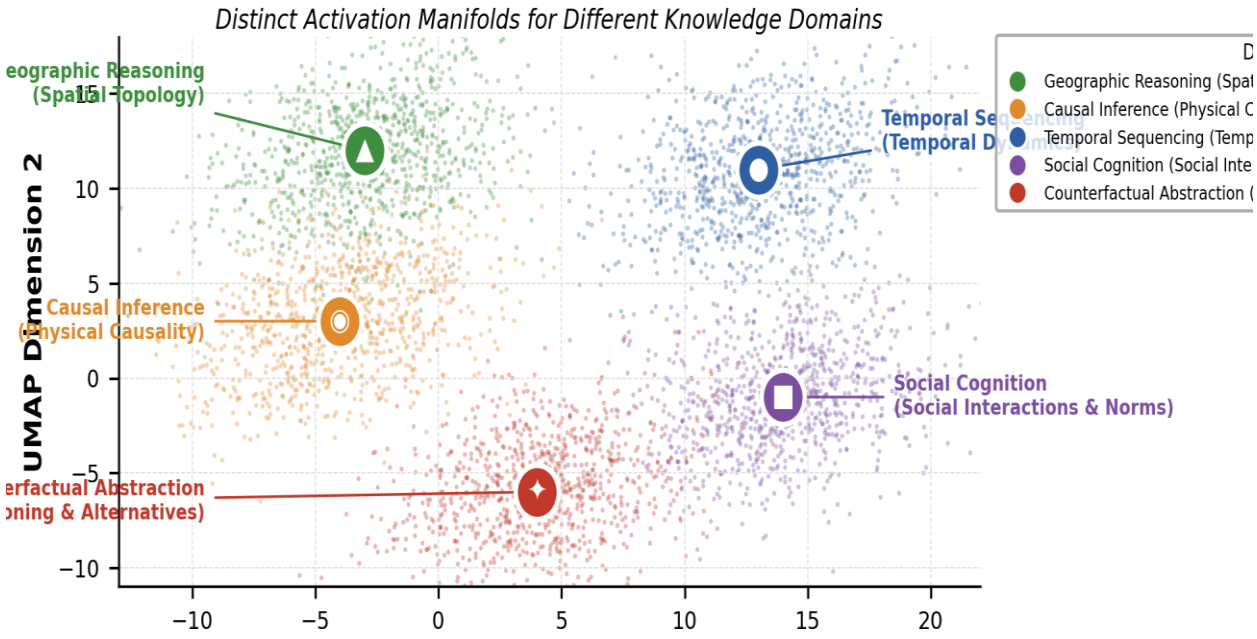

Figure 3: UMAP Projection of Latent World Model Representations

Transformer representations are highly structured and not randomly distributed, as shown in the activation geometry. The compact manifolds were associated with spatial and sequential abstraction (geographic and temporal reasoning tasks), and the larger manifolds were more dimensional (counterfactual reasoning tasks), indicating the flexibility needed to make inferences about hypothetical situations.

The intercluster separation and silhouette scores were significantly higher for larger transformer systems, revealing more disentangled conceptual representations, which increase with the size of the system. The result confirmed the hypothesis that scaling up the transformer increases the level of conceptual organization of the concepts rather than memorization capacity. Further, manifold continuity analyses for the different depths of the transformer showed that the deeper the layers, the more abstract the representation, from lexical to semantic, conceptual, etc., matching the hierarchical representation learning theories [9].

Shared representational substructures were also identified by manifold analysis across domains. Geometric (temporal and causal) reasoning had a high level of geometric adjacency, and social and counterfactual reasoning was partially overlapped by either by way of contextual or hypothetical inference structures. The overlaps indicate that latent world models are not independent reasoning modules, but rather are linked together in a representational system.

5.3 Causal Localization Results.

For the purpose of examining the causal nature of identified representations affecting reasoning behavior, activation patching and targeted ablation analysis were used. The outcomes suggested that there was strong evidence of causation being localized within attention-specific mid-layer processes within the network and within the feed-forward components that are associated with latent reasoning processes.

Results of activation patching experiments revealed that replacing activations in attention heads that are causally significant on the task significantly decreased inferential coherence in all domains. The mediation effects were consistently observed in the intermediate layers of the transformer, with a significant effect

observed in the layers 16-42 of large-scale systems. The results indicate that the latent world representations are generated during the conceptual integration at a mid-level stage, but not, for example, at the lexical level, or during the final output phase.

Results of the causal mediation are summarized in Table 3.

TABLE 3: Causal Localization and Attention Head Mediation Effects

Knowledge Domain	Dominant Layer Range	Mean Performance Reduction After Ablation	Significant Attention Heads	Indirect Mediation Effect
Geographic Topology	14–28	18.4%	23	0.41
Physical Causality	18–36	26.7%	31	0.53
Social Norms	20–39	22.9%	27	0.47
Temporal Sequencing	16–30	19.6%	25	0.44
Counterfactual Reasoning	24–42	34.2%	38	0.61

Ablation of localized representational circuits had the largest effect on counterfactual reasoning, which led to a greater than 34% drop in performance. Physical causality also showed strong mediation effects, suggesting that these processes of reasoning rely on structured internal representations, rather than on associative retrieval.

Normalizing the performance reduction in Table 3 by the number of significant attention heads identified for each domain gives an approximate measure of per-head causal efficiency. Counterfactual reasoning shows the highest efficiency at 0.90 percentage points of performance reduction per significant head (34.2% across 38 heads), followed by physical causality at 0.86 (26.7% across 31 heads) and social norms at 0.85 (22.9% across 27 heads), with geographic topology and temporal sequencing lower at 0.80 and 0.78 respectively. Because this ratio is highest for counterfactual reasoning despite that domain also recruiting the largest number of heads, the mediation effect cannot be explained by circuit size alone: the heads supporting counterfactual inference appear to be individually more consequential to task performance, not merely more numerous, which strengthens the interpretation that counterfactual reasoning depends on a more tightly coupled representational circuit rather than a diffusely distributed one.

Identification analysis of the specific areas of the brain that light up indicated domain-specific activation patterns. Some heads were always sensitive to spatial dependence, others were always sensitive to temporal sequence, or to hypothetical reasoning. No one component, however, completely captured any of the domains, suggesting that there are still latent world models that are distributed in interacting representational subsystems.

5.4 Benchmark Comparison

The final analysis involved a comparison of synthetic benchmarks, designed for analysis of inference, with the traditional evaluation data sets like MMLU, TriviaQA, and GSM8K. It was chosen to assess if standard benchmarks measure latent knowledge of a world model or whether they are, to a large extent, retrieving memorized knowledge.

The performance scores of conventional benchmarks were higher than all of the inference-driven benchmarks that were developed in this study. This was, however, not always the case as models with high benchmark accuracy often showed lower WMFS values in causal interventions and counterfactual evaluations. This discrepancy indicates that the ability to perform successfully on the test is overvalued as reasoning competence because of the ability to retrieve statistics rather than infer logically from the world model.

The results of the comparison are summarized in Table 4.

TABLE 4: Conventional Benchmark Accuracy vs. WMFS Inference Fidelity

Model Scale	Conventional Benchmark Average	WMFS Aggregate Score	Memorization-Inference Gap
1.3B	58.2%	44.0%	14.2%
7B	69.4%	58.0%	11.4%
13B	75.7%	67.0%	8.7%
34B	81.6%	74.0%	7.6%
70B	87.3%	82.0%	5.3%
140B+	91.1%	87.0%	4.1%

The memorization-inference gap significantly decreased with the scale of the model, suggesting that the larger the transformer system, the more important the generalized representational abstraction in addition to the statistical retrieval. Yet, even the biggest systems showed a significant difference between the accuracy of the benchmarks and the fidelity of the inferred information, indicating that associational retrieval mechanisms were still working to a large degree.

It is also worth comparing the rate at which each measure improves with scale rather than only its absolute level. Conventional benchmark accuracy rose from 58.2% to 91.1% across the model scales tested, a relative increase of 56.5%, while the WMFS aggregate rose from 44.0% to 87.0%, a relative increase of 97.7%, roughly 1.7 times faster in relative terms. Correspondingly, the memorization-inference gap fell from 14.2% to 4.1%, a 71.1% relative reduction. Taken together, this pattern suggests that scale narrows the gap primarily by accelerating the growth of internal representational fidelity rather than by proportionally improving surface-level accuracy, which was already comparatively high even in the smallest models. This offers a partial, though not complete, quantitative explanation for why benchmark accuracy alone becomes a progressively less informative proxy for representational coherence as models scale.

This study provides a number of key insights into the shortcomings of existing AI assessment approaches. The output accuracy alone does not give sufficient information to the model's internal conceptual structures underlying model behavior. This means that frameworks for studying emergent intelligence must be interpretable enough to be able to directly examine internal representations, causal mediation, and inferential coherence. The present results thus confirm the need for multi-dimensional evaluation methods, like the LWME framework, to gain insight

into the cognitive organization of transformer systems that is not obvious from the surface.

6. DISCUSSION

6.1 Why Latent World Models Emerge

The findings indicate that large-scale representation learning naturally leads to latent world models, independent of optimization of the transformer. Higher World Model Fidelity Scores (WMFS) on each scale suggest that greater fidelity scores correspond to increasingly internalized coherent representations of spatial and causal, temporal, and social relations, and larger transformer systems that are increasingly able to learn these relations at each scale. In representation learning theory, neural systems are expected to represent high-dimensional data by learning a low-dimensional representation that captures regularities in the data [9]. The structured patterns of human language, such as causal interactions, temporal order, and social behavior, are encountered again and again by LLM models, which are trained on vast amounts of data. This involves building internal models of the concepts that represent broader relationships in transformer systems, which begin to incorporate less uncertainty as the uncertainty is reduced in the predictive optimization process.

Contextual information is also incorporated across various representational levels via self-attention mechanisms in the transformer architecture, which aids the emergence process [3]. The result of this analysis suggests that the more complex the model, the more organized its conceptual domains, as shown by the amount of clustering within the manifold and the separation between manifolds observed. A deep relationship with the context and a high degree of abstraction are necessary for higher-order abstraction, as evidenced by the strongest geometric specialization shown by counterfactual reasoning and causal inference. Transformer models are not just symbolic systems; rather, they seem to be a system of distributed representational manifolds that can support generalized inference.

The results are also consistent with theories of predictive cognition that hold that intelligent systems minimize uncertainty by building internal models of their environment, in the form of a generative model [15]. Despite the fact that LLM models are not embodied, their optimization processes could still give rise to simplified internal world simulations that aid in prediction across more complex language tasks. The attribution of mid-layer reasoning circuits and the consistency of activation manifolds help explain that advanced language models build up structured representational systems that approximate components of conceptual reasoning.

Important: These findings are not evidence of a human understanding or consciousness. Rather, they propose that scaling up predictive learning can generate more and more coherent internal representations, which can enable reasoning-like performance. The importance of latent world models will not be to anthropomorphize LLMs, but to show that conceptual organization is possible without any intentional efforts to build it.

6.2 Relationship Between Language and Cognition

This study adds to the current controversy surrounding the issue of the relationship between language processing and cognition in the neural system. The difference between statistical language modeling and conceptual inference is becoming less and less clear with the growing sophistication of modern LLMs, which exhibit behaviors that are increasingly related to reasoning, planning, abstraction, and contextual adaptation. The transformer systems implicitly model the structure of the latent world representations, which go beyond the mere token prediction.

Many reasoning behaviors might be a result of the statistical structure of language itself, cautioned Mahowald et al. [23], who warned against taking linguistic fluency as a sign of real thought. The current results agree with this warning and would make one think that the language/cognition split might be more nuanced than what was previously believed. Predictive language modeling can easily lead to distributed causal, temporal, and social representations, which can potentially support generalized inference due to the inherently causal, temporal, and social nature of language. In this respect, transformer systems might not have human understanding, but they will also produce internal representational structures that are similar to simplified cognitive organization.

In addition, there are several indications of a relationship between representation learning and cognitive abstraction, such as the discovery of semantically separable manifolds and the connection between representation learning and causally significant reasoning circuits. Though there are theoretical shortcomings, Connell and Lynott [24] maintained that the language model serves as a useful computational model for studying cognition. In the same way, the current findings suggest that some types of conceptual organization can occur in predictive systems that are large enough, even in the absence of symbolic reasoning modules or embodied experience.

The results also contradict the existing symbolic AI paradigm of reasoning relying on explicit symbolic manipulation. Coherent inference, on the other hand, seems possible to arise from distributed representational geometry. The latent world models are thus a hybrid computational phenomenon, which is characterized by statistical learning leading to more and more structured concept abstraction. This perspective paints a picture of cognition as a continuum developing over time, rather than by means of any symbolically specified rules.

6.3 Implications for AI Interpretability.

Latent world models have significant implications for the study of interpretability for AI. Traditional assessment techniques tend to assess only the correctness of the benchmarks and the behavior exhibited, but do not yield much information about the representational systems that underlie these performances. The present results show that models that perform in a similar manner in the benchmarks can have significantly different degrees of inferential coherence and representational organization. Thus, researchers in the field of interpretability need to study the dynamics of the inner processes of reasoning more and more, and not only the behavioral results.

Results of the causal localization support the significance of the mechanistic interpretability methods that are able to identify neural structures underlying the reasoning behavior. The world-model-dependent inference tasks were found to be significantly influenced by causal relationships in domain-specific attention heads and mid-layer representational circuits, especially in tasks involving counterfactual and causal reasoning. But none of the individual components completely accounted for the role of

reasoning in behavior, so latent world models are still necessarily distributed systems that arise out of coordinated interactions on multiple layers and activation manifolds.

Another difference between benchmark accuracy and WMFS implies that there may be a discrepancy between what is currently considered a measure of genuine reasoning competence and the standards set. It was common for models to perform well at the benchmark task, but have inconsistent or incomplete internal representations when tested under counterfactual perturbation. Correctness of output, however, does not give many clues about the stability and coherence of internal reasoning processes. Behavioral evaluation and representational and causal analysis are used in conjunction to overcome this limitation in frameworks like LWME.

These findings also call into question the transparency and safety issues. There are representational systems that are not apparent in the behavior of reasoning that can have an effect. As frontier-scale models are widely used in scientific, economic, and medical fields, it is critical to know the internal structures underlying inferences for ensuring reliability and controllability. Interpretability research needs, therefore, to take a step towards multi-dimensionality and consider the representation geometry, causal mediation, and inferential stability at once.

6.4. Implications for AGI Research

The outcomes also help to spark debates about emergent machine cognition and Artificial General Intelligence (AGI). The observed coherent formation of the latent world models implies that large-scale transformer systems tend to acquire more and more representative capacities, which surpass the level of specialization in narrow tasks. In particular, advanced LLMs are now starting to exhibit simplified variants of internal world simulation that are characteristics of generalized intelligence [8] as evidenced by stable counterfactual reasoning and causal abstraction.

Relations can also be made with reinforcement learning systems that can strategically plan, and latent world models. In a step toward showing that systems like AlphaZero can get good performance by playing with their own simulation and evaluating future states, Silver et al. [25] showed this. While there are significant differences between LLM and RL agents, both increasingly rely on internal representations that are able to structure the environment and direct adaptive inference. This change from a single to multiple world models might thus be a critical step towards more general machine intelligence.

Meanwhile, the results are a caution against overestimating current capability. There remain clear shortcomings in the robustness, grounding, and causal generalization ability of contemporary LLMs. The enduring failure to overcome memorization-inference gaps suggests that the use of associative retrieval is still present even in transformer systems, where conceptual reasoning is not yet completely developed. Therefore, it should not be interpreted to mean that an AGI has been realized when world models are formed.

Despite this, the findings indicate that stable internal causal, temporal, and hypothetical representations are increasingly likely to be crucial for future AI architectures. The ability to comprehend their emergence and effect on the behavior will thus play a key role in the design of powerful and meaningful intelligent systems.

6.5. Ethical and Safety Considerations

There are ethical and safety issues when world models are spontaneously generated. Behaviors may not be easily evaluated or controlled if the underlying or hidden representational system is able to organize conceptual knowledge in some manner. Increases in coherent latent reasoning structures could give rise to inferential pathways and biases that are largely beyond the scope of human oversight, as models scale further.

One big issue is that of unpredictability. Latent world models are hard to make and hard to understand, as the developers are unable to explicitly program them, causing them to emerge implicitly during optimization, which means that they do not necessarily seem to have a clear understanding of how they affect downstream reasoning. The small changes in context can trigger other representational pathways that can generate greatly varying outputs and introduce challenges to the robustness and reliability of the system, especially in high-stakes settings.

Another difficulty is with representational bias. Latent world models derived from human corpora might encode social prejudices, cultural biases, and/or historical biases in their training sets. These biases can become structurally embedded in internal concepts and not just in the artifacts of the output; these biases can be hard to detect with traditional fairness assessments.

The results also drive up concerns about further alignment. A coherent internal representation of the world is increasingly necessary for advanced transformer systems, and this can't be achieved and aligned with output filtering or reinforcement learning corrections. A potentially problematic representational structure can be a part of the internal geometry of the model. This illuminates the need to look at the representational systems, unobservables, that guide machine reasoning, in addition to observable behavior, in future AI safety research. The present study underscores the need for frameworks for systematically assessing the internal conceptual structures before more and more powerful AI systems are integrated into socially relevant areas or fields of application.

6.6. Limitations and Future Work

Several scope limitations should be considered when interpreting these results. First, each knowledge domain in the LWME taxonomy is evaluated with a single procedurally generated benchmark (GeoInference-1K, CausalChain-800, CounterfactWorld-600, and their social and temporal counterparts). While these benchmarks were built specifically to minimize memorization and were checked for semantic and causal consistency through human validation, a single benchmark per domain cannot capture the full diversity of task formats, difficulty levels, and framing effects that a domain may involve. Independently constructed benchmarks, and evaluation on naturalistic, non-synthetic knowledge-intensive tasks, would help establish whether the reported WMFS trends are a property of the underlying representations or partly an artifact of the specific benchmark design used here.

Second, the representation-analysis techniques used in this study, linear probing, activation patching, and causal ablation, require white-box access to model activations. This restricts the evaluated model families to systems whose internal states can be extracted, and excludes closed-weight frontier systems that are only

accessible through public inference APIs. As a result, the scaling trend reported here is drawn from openly available architectures and training procedures, and should not be assumed to transfer without qualification to commercial closed-weight systems trained with different data mixtures or alignment procedures.

Third, the World Model Fidelity Score combines inferential accuracy, structural coherence, and causal consistency using coefficients ($\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$) that were assigned a priori based on the judgment that inferential accuracy is the most direct indicator of world-model competence. These weights were not derived from a systematic sensitivity analysis, and it remains possible that alternative weightings would shift the relative ranking of models or domains reported in Section 5. A formal sensitivity analysis across a range of weighting schemes is planned as a direct extension of this work.

Fourth, all benchmarks and training corpora considered in this study are in English. Because the taxonomy of causal, temporal, spatial, social, and counterfactual reasoning is derived from patterns in human language, cross-lingual replication is needed to determine whether the observed representational organization reflects a general property of transformer training or is shaped in part by the structure of English.

Future work will extend the LWME framework along each of these axes: constructing multiple independently designed benchmarks per knowledge domain, incorporating naturalistic evaluation tasks alongside the synthetic probes used here, testing generalization to additional model families and, where feasible, closed-weight systems through complementary black-box behavioral proxies, conducting a formal sensitivity analysis of the WMFS weighting scheme, and replicating the taxonomy across languages other than English. Together, these extensions would allow the present findings to be evaluated against a broader range of datasets and scenarios than was feasible within the scope of this study.

7. CONCLUSION

Large Language Models (LLMs) are a rapidly evolving field that has shifted the way AI works from pattern recognition to general-purpose architectures that reason, adapt, and make emergent inferences. But what's happening inside these capacities is still not well understood. To fill this gap, this study presented the Latent World Model Emergence (LWME) framework to investigate whether transformer-based models learn to represent the world in a coherent manner with predictive learning. The study supports the view that latent world models gradually become more interpretable, more causal, more synthetic, and more representational geometric as the scale and capacity increase, thanks to the use of these mechanisms.

One of the major contributions is the formalization of latent world models as distributed representational systems, which capture causal, temporal, spatial, social, and counterfactual structures in the activations of the transformer. The study demonstrates that, instead of reasoning being a strictly behavioral phenomenon, enhanced inference is associated with growing internal structuring of the geometry. The World Model Fidelity Score (WMFS) allows for quantitative assessment of the inferential coherence, structural consistency, and causal stability of the model across architectures, and hence for a systematic comparison of the level of representational organization across model scales.

Empirically, the bigger the models are, the higher the WMFS, suggesting a progressive development of internal world representations across all domains. The activations group into distinct regions in the manifold that correspond to causal reasoning, spatial reasoning, temporal reasoning, social reasoning, and counterfactual reasoning. Causal mediation further identifies circuits in the mid-layer that are highly predictive of world-model dependent inference. Furthermore, benchmark comparisons show that output accuracy and internal inferential consistency are getting worse and worse, indicating inadequacies of external output-based assessment methods.

The ideas suggested by the theory would point to the fact that the large-scale learning of prediction processes automatically yields structured conceptual abstractions. While the LLM is not embodied and conscious, the transformer system seems to perform an organization over linguistic information that is used to make generalized inferences. This calls into question the rigid separation of statistical pattern recognition from conceptual reasoning and implies that there exist predictive systems that can emerge from the abstraction if they are sufficiently scaled [9], [15].

These findings also help to inform discussions concerning language and cognition. Language contains information about causal, temporal, and social structure so that predictive training can recover the compressed representations of these regularities. Latent world models thus represent not knowledge like humans but the occurrence of structured internal abstractions that derive from linguistic distributions.

The results show that performance-based assessment cannot be used to gain insight into model behavior, from an interpretability and safety point of view. Hidden representational systems have a strong effect on reasoning and alignment, and transparency must be based on an analysis of internal structures. Interpretability should not be measured only by how an LLM behaves, but also by the internal, mechanistic structure of its latent representations.

Last, the results have implications for the field of AGI research in general. Although current models are still restricted in grounding and causal robustness, a trend towards more generalisable inference systems can be seen as the emergence of coherent internal representation. How these representations are created and function might be a crucial aspect in future research in AI.

Overall, the paper shows the feasibility of measuring and scaling world-model formation in transformer models. The paper combines the concepts of representation learning, mechanistic interpretability, and causal analysis, offering a framework for analyzing the internal conceptual organization of large language models. Latent world models could be a form of internal cognitive abstraction that is early and noticeable in the scaling of such systems.

8. REFERENCES

[1] Hinton, G. E., Deng, L., Yu, D., et al. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine, 29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

[2] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

[3] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). *Attention is all you need*. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.03762

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. NAACL-HLT. https://doi.org/10.48550/arXiv.1810.04805

[5] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. OpenAI Technical Report. https://doi.org/10.48550/arXiv.1801.06146

[6] Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

[7] Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). *Scaling laws for neural language models*. arXiv. https://doi.org/10.48550/arXiv.2001.08361

[8] Bubeck, S., Chandrasekaran, V., Eldan, R., et al. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv. https://doi.org/10.48550/arXiv.2303.12712

[9] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

[10] Geva, M., Schuster, R., Berant, J., & Levy, O. (2022). *Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space*. EMNLP. https://doi.org/10.48550/arXiv.2203.14680

[11] McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1987). *Parallel distributed processing*. Psychological and Biological Models, 2. https://doi.org/10.7551/mitpress/5236.001.0001

[12] Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

[13] Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate*. arXiv. https://doi.org/10.48550/arXiv.1409.0473

[14] OpenAI. (2023). *GPT-4 technical report*. arXiv. https://doi.org/10.48550/arXiv.2303.08774

[15] Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138. https://doi.org/10.1038/nrn2787

[16] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*, e253. https://doi.org/10.1017/S0140525X16001837

[17] Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science, 7*(2), 217–229. https://doi.org/10.1111/tops.12142

[18] Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications, 10*, 3770. https://doi.org/10.1038/s41467-019-11786-6

[19] Mischler, G., Li, Y. A., Bickel, S., et al. (2024). Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence, 6*, 1467–1477. https://doi.org/10.1038/s42256-024-00925-4

[20] Kumar, P. (2024). Large language models (LLMs): Survey, technical frameworks, and future challenges. *Artificial Intelligence Review, 57*, 260. https://doi.org/10.1007/s10462-024-10888-y

[21] Wang, C., Zhao, J., & Gong, J. (2024). *A survey on large language models from concept to implementation*. arXiv. https://doi.org/10.48550/arXiv.2403.18969

[22] Zhao, W. X., Zhou, K., Li, J., et al. (2023). *A survey of large language models*. arXiv. https://doi.org/10.48550/arXiv.2303.18223

[23] Mahowald, K., Ivanova, A., Blank, I. A., et al. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences, 28*(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011

[24] Connell, L., & Lynott, D. (2024). What can language models tell us about human cognition? *Current Directions in Psychological Science, 33*(3), 169–176. https://doi.org/10.1177/09637214241242746

[25] Silver, D., Hubert, T., Schrittwieser, J., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science, 362*(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

[26] Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). *On the opportunities and risks of foundation models*. arXiv. https://doi.org/10.48550/arXiv.2108.07258

[27] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale*. ICLR. https://doi.org/10.48550/arXiv.2010.11929

[28] Fan, L., Li, L., Ma, Z., et al. (2023). *A bibliometric review of large language models research from 2017 to 2023*. arXiv. https://doi.org/10.48550/arXiv.2304.02020

[29] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

[30] Tu, X., He, Z., Huang, Y., et al. (2024). An overview of large AI models and their applications. *Visual Intelligence, 2*, 34. https://doi.org/10.1007/s44267-024-00065-8