CogEvolve: A Multimodal Benchmark for Evaluating Relational Reasoning in Semantic Extension

Jingjie Zeng[1]*, Huayang Li[1]*, Liang Yang[1,2]†, Yuanyuan Sun[1], Shaowu Zhang[1], Hongfei Lin[1]
[1]School of Computer Science and Technology, Dalian University of Technology, China
[2]Key Laboratory of Social Computing and Cognitive Intelligence, Ministry of Education, China
jjtail@mail.dlut.edu.cn, liang@dlut.edu.cn

Abstract

Human cognition excels at extending knowledge through analogy, where word meanings evolve along structured pathways from concrete prototypes to abstract senses via metaphor and metonymy. **Do Large Language Models (LLMs) internalize this generative logic, or merely mimic statistical patterns?** To investigate this, we introduce **CogEvolve**, a cognitive linguistic benchmark designed to test these evolutionary pathways across textual and visual modalities. Our evaluation reveals a distinct cognitive profile: models function as "Super-Associators" expert at static recognition yet fail at causal reasoning. In text, they exhibit a Frequency-Primacy Conflation, confusing statistical prevalence with cognitive basicness. Crucially, this reasoning collapses further in the visual domain. We term this deficit the Ungrounded Arrow: models possess high-fidelity concept representations (the "dots") but lack the transformational operators (the "arrows") essential for true relational understanding[1].

1 Introduction

A cornerstone of human intelligence is the ability to extend knowledge via analogy, mapping familiar concepts onto novel domains (Gentner, 1983). In language, this manifests as semantic extension: meanings evolve along structured pathways—from concrete prototypes to abstract senses via metaphor and metonymy—governed by cognitive principles (Rosch, 1975; Lakoff and Johnson, 2008; Kennington and Natouf, 2022).

However, as illustrated in Figure 1, a gap exists between human embodied logic and machine statistical learning. For humans, "crust" is grounded in the physical experience of bread before extending to geology; for LLMs, the scientific term "Earth's crust" often statistically overshadows the mundane root due to reporting bias. This raises a fundamental question: **Do models merely mimic the statistical surface of language (the "dots"), or have they internalized the generative logic that connects them (the "arrow")?** While benchmarks like WSD (Bevilacqua et al., 2021) and VQA (Antol et al., 2015) evaluate static association, they fail to probe this directional causality.

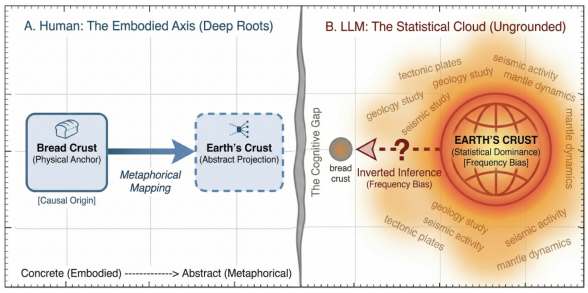

Figure 1: **The Conceptual Topography of Meaning Grounding.** (A) **Human** follows a directed causal path: abstract concepts (e.g., "Earth's Crust") are anchored in and projected from concrete embodied experiences ("Bread Crust"). (B) **LLM** operates in a "Statistical Cloud". Driven by reporting bias in corpora, high-frequency derived senses overwhelm the mundane cognitive root, creating a gravitational pull that inverts the inference direction.

We hypothesize that current architectures, optimized for next-token prediction, excel at learning co-occurrence patterns (association) but may not inherently acquire the abstract operators required for directional inference (reasoning). To test this, we introduce **CogEvolve**, a diagnostic benchmark grounded in cognitive linguistics. As illustrated in Figure 2, our framework models semantic evolution across three distinct layers—Symbolic, Structural, and Schematic. Crucially, unlike previous datasets relying on raster images, CogEvolve models visual concepts using Scalable Vector Graphics (SVGs) at the Schematic layer. As code-based represen-

*Equal contribution.
†Corresponding Author
[1]The code and datasets are available at https://github.com/jjtail/CogEvolve-ACL2026

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25943–25960

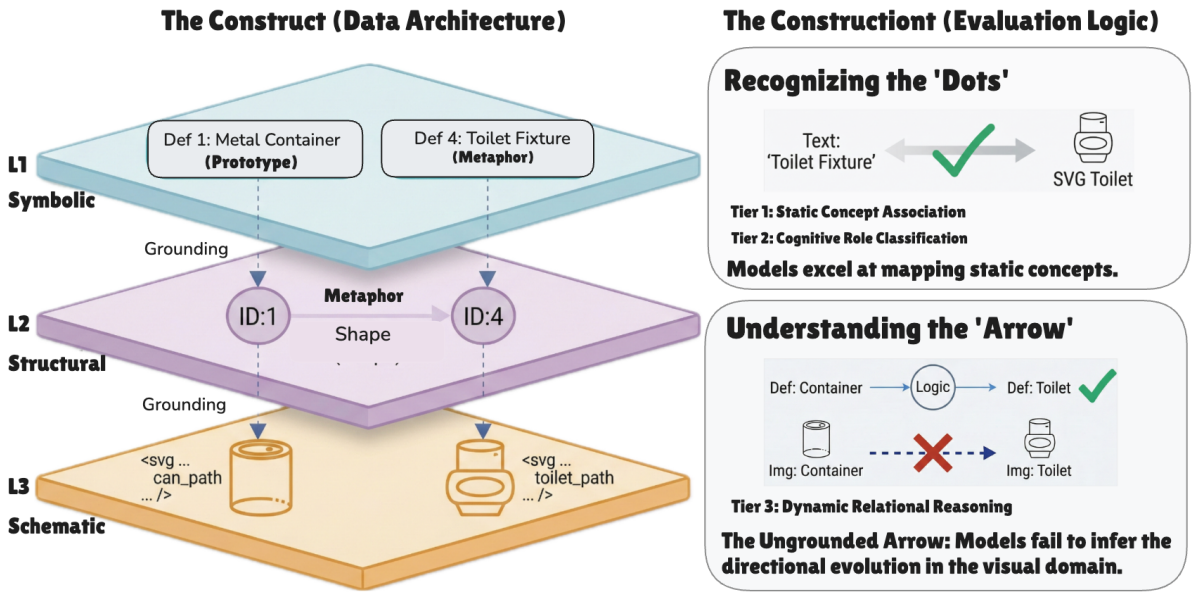

Figure 2: **The Conceptual Framework of CogEvolve. (Left) Data Architecture:** We model polysemous words across three aligned layers. The **Symbolic Layer (L1)** provides textual definitions; the **Structural Layer (L2)** defines the causal history as a directed graph (e.g., Prototype $\rightarrow$ Metaphor); and the **Schematic Layer (L3)** grounds these concepts in Structured SVGs to function as a visual Chain-of-Thought. **(Right) Evaluation Logic:** The evaluation forms a "funnel" tiered by cognitive complexity. **Tier 1 & 2** assess static competence (Recognizing the "Dots"), verifying that models can perceive and classify concepts. **Tier 3** tests dynamic relational reasoning (Understanding the "Arrow"), requiring models to reconstruct the generative path of semantic extension.

tations, SVGs decouple schematic structure from perceptual noise, explicitly encoding the geometric transformations inherent in semantic change. This design allows SVGs to function as a Visual Chain-of-Thought (CoT)—a transparent intermediate state that exposes the logical topology of visual reasoning, bridging the gap between symbolic definitions and pixel-level perception.

Our evaluation reveals a divergent evolutionary pattern in current AI systems. In the textual domain, frontier models have successfully internalized the generative logic of meaning, manifesting as a **"Symbolic Arrow"** that achieves pairwise inference accuracy comparable to human baselines. Control experiments confirm this capability is driven by abstract semantic operators rather than lexical memorization. Conversely, this reasoning capability collapses in the multimodal domain. We identify an **"Ungrounded Arrow"** where models, despite exhibiting expert-level proficiency in static recognition, struggle to map abstract visual transformations to logical sequences. They effectively perceive the conceptual nodes but fail to infer the directional vector of change, exposing a fundamental deficit in grounded relational reasoning.

In light of this dichotomy between symbolic competence and grounded disconnect, our work advances the understanding of machine cognition through three primary contributions:

• **A Cognitively-Grounded Benchmark:** We introduce **CogEvolve**, the first dataset explicitly modeling the generative logic of semantic evolution. By utilizing Structured SVGs as interpretable intermediate representations, we provide a testbed that decouples structural reasoning from perceptual noise, achieving high inter-annotator agreement ($\kappa = 0.82$).

• **Systematic Diagnosis of Reasoning Boundaries:** We provide a quantitative mapping of the disparity between symbolic and grounded reasoning. Our results indicate that while the reasoning gap has closed in text, a substantial deficit remains in schematic visual reasoning.

• **The "Ungrounded Arrow" Hypothesis:** We formalize the mechanism behind this deficit. We identify that while models excel at static cross-modal association, they fail at dynamic relational reasoning. This suggests current architectures lack a unified cognitive space to process generative logic across symbolic and schematic representations.

2 Related Work

Cognitive Foundations of Meaning and Analogy. Human cognition extends knowledge dynamically rather than statically storing facts. Central to this process is Prototype Theory (Rosch, 1975), where meanings evolve from a concrete core via analogy. This extension is formalized by Structure-Mapping Theory (Gentner, 1983) in psychology and Conceptual Metaphor Theory (Lakoff and Johnson, 2008) in linguistics, which ground abstract thought in embodied experience. While metonymy extends meaning via contiguity (Barcelona, 2012), current computational benchmarks rarely evaluate whether models internalize these structural principles of semantic derivation.

These theories collectively paint a picture of semantic knowledge as a structured, dynamic, and generative system. Our benchmark, CogEvolve, is a direct attempt to computationalize this view, creating a testbed to evaluate whether AI models exhibit similar structural principles.

Probing the Cognitive Plausibility of LLMs. The question of whether LLMs are mere "stochastic parrots" (Bender et al., 2021) or are developing more genuine forms of understanding is a central debate. Researchers have approached this by probing LLMs for human-like cognitive biases and abilities(Hardy et al., 2023; Strachan et al., 2024). Studies have investigated their capacity for Theory of Mind (Kosinski, 2023), physics (Xu et al., 2024), and moral reasoning (Ji et al., 2025). In linguistics, this has involved testing their grasp of syntactic structures (Linzen et al., 2016) and semantic roles. Furthermore, recent works have begun to evaluate physical concept understanding (Yu et al., 2025) and metaphor processing through analogies (Boisson et al., 2024; Tong et al., 2024).Our work fills this gap by focusing on the diachronic logic of meaning change, testing a faculty that is fundamental to generative intelligence: the ability to extend familiar concepts to novel domains via structured derivation.

Symbol Grounding and Schematic Abstraction. Traditional multimodal benchmarks (VQA) primarily rely on pixel-based grounding. However, recent work like VGBench (Zou et al., 2024) challenges this paradigm, demonstrating the efficacy of vector graphics (VG) as a concise, code-based visual representation. Building on this shift, we probe a deeper layer: **image schemas** (Johnson, 2013)—recurring patterns (e.g., CONTAINER) that structure thought. Unlike VGBench's focus on visual understanding, CogEvolve leverages the structural transparency of SVGs to test **schematic reasoning**: challenging models to recognize analogical transformations (e.g., a container becoming a boundary) rather than merely rendering shapes.

By doing so, we address a critical limitation in current evaluations: moving beyond the static *perception* of visual concepts to probing the dynamic *causality* of their semantic grounding.

3 The CogEvolve Benchmark and Task Definitions

To empirically test the hypotheses raised in our introduction, particularly the distinction between associative and relational capabilities in LLMs, we develop **CogEvolve**. This section details the design principles of our benchmark.

3.1 Conceptual Definitions

To ensure conceptual clarity and avoid terminology confusion, we define the interdisciplinary terms central to our framework. **Cognitive Basis** relates to the human psychological processes—informed by Prototype Theory and Conceptual Metaphor Theory—that govern how knowledge is dynamically extended. **Evolutionary Stages** denote the synchronic semantic extension of a word: a logical hierarchy showing how meaning extends from a concrete source to an abstract target, capturing structural derivation rather than historical etymology. Finally, **Schematic Grounding** in our context refers to anchoring abstract semantic relationships into computable schemas (SVGs). While this does not fulfill true physical, situated symbol grounding in a concrete sensory environment, it acts as an interpretable "Visual Chain-of-Thought".

3.2 Motivation for Benchmark Design

The cornerstone of our framework is **CogEvolve**, a benchmark designed to evaluate the *generative logic* of semantic extension, rather than mere vocabulary coverage. Unlike traditional WSD datasets that model polysemy as a flat list of isolated senses, CogEvolve represents meaning as a directed acyclic graph, encoding the causal derivation from embodied roots to abstract extensions.

From Static Classification to Generative Inference. Standard benchmarks evaluate whether

models can *recognize* a meaning in context. However, they fail to probe whether models understand *why* a word holds multiple meanings. We argue that true semantic understanding requires modeling the **evolutionary vector**—the directed path from a **Prototype** (e.g., a physical container) to its derived senses via **Metaphor** and **Metonymy**. CogEvolve captures this structure, forcing models to distinguish between *statistical correlation* (co-occurrence) and *cognitive derivation* (causality), a distinction that cannot be learned solely from distributional patterns in text.

Diagnostic Motivation: Disentangling Reasoning from Retrieval. Ultimately, CogEvolve is constructed to serve as a cognitive diagnostic. In current evaluations, models can achieve high performance merely by memorizing surface-level correlations, masking their inability to grasp the underlying causal logic. By enforcing a structural separation between *embodied derivation* and *statistical frequency*, this benchmark provides the necessary ground truth to rigorously determine whether models are merely retrieving facts or simulating the generative processes of human thought.

3.3 Benchmark Construction

Human-in-the-Loop Data Construction Protocol. To achieve this structural validity while maintaining scale, we employed a rigorous Human-Architected, AI-Assisted Protocol. The creation process leverages human experts for cognitive ground truth while utilizing Large Language Models as structured processors.

(1) Knowledge-Based Sense Consolidation. We first source a candidate pool of 6,431 polysemous words from WordNet (Miller, 1995) and their relational data from ConceptNet (Speer et al., 2017). This raw information is then processed by a diverse set of LLMs. Their task is not reasoning, but structured data processing: consolidating duplicate definitions, paraphrasing them for clarity, and organizing them into a consistent format for human annotation.

(2) Human-Driven Cognitive Annotation. With the structured sense sets as input, our team of trained human annotators performed the core cognitive tasks. Their first step is to definitively identify the cognitive prototype for each word. We acknowledge that an individual's cognitive prototype can be influenced by their personal background. Therefore, our annotation guideline is not based on personal intuition, but on the established principles of concreteness and cognitive basicness from Prototype Theory (Rosch, 1975). The goal is to simulate the linguistically-recognized, generative semantic core. Following this, annotators manually construct the directed graph of semantic evolution by creating subpaths.

(3) Visual Schema Generation. For the visual layer, we implement a Hybrid Human-Architected Protocol (detailed in **Appendix E**). For senses, annotators first author visual narratives describing the abstract essence of a concept (e.g., a container expanding to become a limit). Unlike raster images, SVGs are code-based and structural. We utilize a code-generation model as a syntactic compiler to render these narratives into SVG code. By enforcing structural constraints, we ensure that visual metaphors are mathematically identical across evolutionary stages. Human analogical reasoning relies on abstraction—stripping away sensory details (texture, lighting) to reveal underlying structural invariants. By using SVGs, CogEvolve mimics this process of cognitive compression. The vector format forces the visual representation to operate at the level of schematic logic rather than pixel perception, ensuring that the benchmark evaluates the model's ability to reason about geometric essence—the true medium of conceptual metaphor—rather than its sensitivity to photorealistic noise.

The resulting benchmark comprises 6,431 polysemous words. A key feature of **CogEvolve** is its structural complexity: while approximately 64% of subpaths represent direct binary extensions (Prototype $\rightarrow$ Extension), over 36% involve multi-step evolutionary chains ($N \geq 3$). For instance, the semantic evolution of "cloud" progresses sequentially from physical density $\rightarrow$ obscurity $\rightarrow$ suspicion. This forces models to perform complex transitive reasoning rather than simple direct association. The inter-annotator agreement for the critical path construction task achieved a Cohen's Kappa of $\kappa = 0.82$, demonstrating the high reliability of our human-authored structures.

The Three-Layered Annotation Schema. To capture the richness of this process, each entry in CogEvolve is annotated across three layers, providing a multi-faceted representation of semantic knowledge. We use the word "can" as an illustrative example throughout this section (visualized in Figure 3):

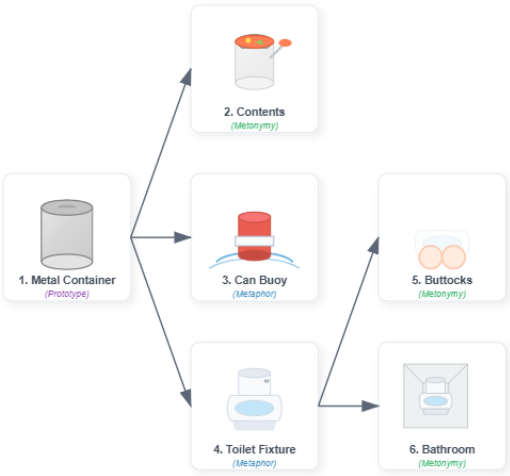

Figure 3: The three-layered representation of "can". Arrows indicate the direction of semantic derivation from prototype along distinct conceptual paths.

(1) Symbolic Layer (Textual Definitions): This layer provides standard textual information, including a concise definition, and its cognitive role for each sense. For "can", this includes the prototype "Metal Container", and extensions such as "Contents" (Metonymy) and "Toilet Fixture" (Metaphor).

(2) Structural Layer (Relational Pathways): This is the logical backbone. It explicitly defines the generative relationships using directed subpaths. For "can", the structure includes paths like `1 -> 2` (container extends to contents). This annotation transforms a flat list of senses into a structured, directed graph.

(3) Schematic Layer (Visual Grounding): To test for deeper, modality-independent understanding, we ground each concept in a visual image schema. These are SVG illustrations designed to capture the core essence of a concept. For "can", the prototype (id 1) is a simple metal can. Its derivations visually encode the analogical transfer: the "Contents" sense (id 2) shows food inside a translucent can, while the "Toilet Fixture" sense (id 4) abstracts the can's shape into a toilet.

3.4 A Tiered Suite of Diagnostic Tasks

Built upon the CogEvolve, we design a suite of tasks that function as a "cognitive diagnostic funnel". This hierarchical structure is engineered to systematically disentangle simple pattern matching from genuine understanding. By increasing cognitive complexity at each stage, the evaluation progressively filters out shallow associative capabilities (the "dots") to isolate the core faculty of generative inference (the "arrow"). The diagnostic process proceeds through three layers:

Tier 1: Static Concept Association. This tier establishes a baseline for a model's ability to link static, isolated concepts. It directly tests the associative capabilities.

- **Task 1a (Textual WSD):** A text-only task to assess contextual association. Given a sentence, the model must select the correct textual definition for a target word.
- **Task 1b (Cross-modal Alignment):** A multimodal task to assess visual-semantic association. The model must match a visual schema to its correct textual definition, and vice-versa.

Tier 2: Cognitive Role Classification. This tier probes whether models understand the structural properties of concepts, moving beyond simple association.

- **Task 2 (P/M/T Classification):** A text-only task where the model must classify a word's usage as a 'Prototype', 'Metaphor', or 'Metonymy', testing its grasp of cognitive-linguistic categories.

Tier 3: Dynamic Relational Reasoning. This core tier directly evaluates the model's ability to infer the directional relationships between concepts—the "arrow". These tasks are designed as a pure test of logical inference. The model is given a shuffled set of concept definitions without any explicit cognitive labels. To succeed, it must deduce the underlying logical progression by analyzing the semantic content of the definitions. This requires recognizing fundamental cognitive principles, such as inferring a path from concrete to abstract, from a general case to a specific instance, or from a whole to its part. This task therefore directly probes a model's capacity for generative reasoning, rather than its knowledge of linguistic terminology.

- **Task 3a (Textual Path Ordering):** A text-only reasoning task. Given a set of textual definitions from a subpath in random order, the model must reconstruct the correct evolutionary sequence.
- **Task 3b (Visual Path Ordering):** The same task, but input consists solely of visual schemas. This tests if model can infer logical "arrow" from visual transformations alone.

3.5 Models and Evaluation Protocol

To ensure a comprehensive evaluation, we select a diverse suite of state-of-the-art models, representing a cross-section of the current LLM and MLLM landscape.

Models. For text-only tasks, our evaluation include leading closed-source models: Claude-4-Sonnet (Anthropic, 2025), Gemini-3-Pro(Gemini, 2025), ChatGPT-5.1 (OpenAI, 2025), and Grok-4-fast(xAI, 2025). We also include prominent open-source models: Deepseek-r1 (Guo et al., 2025), Gemma-3-12B-it (Team et al., 2025), Qwen3-8B (Qwen, 2025)), Llama3.1-8B-Instruct (Meta, 2024), and Mistral-8B-Instruct (Mistral, 2024).

For multimodal tasks, we evaluate leading vision-language models: Gemini-3-pro, Qwen3-VL-8B-Instruct (Bai et al., 2025), ChatGPT-4o-Latest (OpenAI, 2024), and Gemma-3-12B-it.

Evaluation Metrics. For association and classification tasks (Tier 1 & 2), we use Accuracy. For the more complex relational reasoning tasks (Tier 3), we employ a multi-faceted protocol to diagnose, especially for paths of length $N > 2$:

Exact Match Accuracy: A strict, all-or-nothing measure of whether the entire sequence was correctly reconstructed.

Mean Kendall's Tau: A rank correlation coefficient ($\tau \in [-1, 1]$) that measures the overall trend correctness of the sequence.

Mean Pairwise Accuracy: A fine-grained metric measuring the proportion of correctly identified pairwise precedence relationships. A high score coupled with a low exact match score indicates a failure in global sequence construction rather than local comparison.

4 Experiments

Our experiments follow the tiered diagnostic framework to systematically probe the distinction between models' associative capabilities and their relational reasoning skills.

4.1 Human Baselines

To distinguish basic intuition from expert reasoning, we compare models against a **Naive Baseline** ($N = 30$) of undergraduates (testing commonsense semantic intuition) and an **Expert Reference Baseline** ($N = 10$) of cognitive linguistics researchers. Given the limited expert sample size, we treat this group as a reference point for task solvability and qualitative comparison rather than a population-level upper bound. Full protocols and prompt mappings are detailed in Appendix D.

4.2 Finding 1: Models as "Super-Associators"

Our first tier assesses the fundamental ability to link static symbols to referents. As summarized in Table 1, models demonstrate exceptional proficiency, establishing themselves as what we term "Super-Associators".

In Textual WSD (Task 1a), models like *Claude-4-Sonnet* (96.9%) and *GPT-5.1* (95.8%) match or exceed the Expert Human baseline (95.2%), demonstrating mastery over symbolic disambiguation.

Crucially, this proficiency extends to the multimodal domain. In Cross-modal Alignment (Task 1b), *Gemini-3-Pro* surpasses experts with 92.0% accuracy in Image-to-Text recognition and ∼90% in Text-to-Image retrieval.

This bidirectional mastery serves as a vital control condition for our study. It confirms that the visual concepts are correctly indexed in the model's latent space—effectively, the model has successfully acquired the conceptual "dots". This distinction is pivotal: it ensures that the reasoning collapse we later observe in Finding 3 is not a consequence of perceptual blindness, but can be attributed to a deficit in the underlying relational logic.

Task	Model / Group	Score (%)
Task 1a: Textual WSD (Static Association)		
	Claude-4-Sonnet	**96.9**
	Gemini-3-Pro	96.6
	GPT-5.1	95.8
	Expert Human	*95.2*
	Deepseek-r1	93.9
	Naive Human	89.2
	Grok-4-fast	88.6
	Gemma-3-12B-it	80.3
	Qwen3-8B	77.3
	Llama3.1-8B-Ins	70.8
	Mistral-8B-Instruct	68.3
Task 1b: Cross-modal Alignment *(Format: Image-to-Text / Text-to-Image)*		
	Gemini-3-Pro	**92.0 / 90.0**
	Qwen3-VL-8B	91.4 / 89.6
	Expert Human	90.0 / 89.0
	ChatGPT-4o-Latest	89.0 / 88.0
	Gemma-3-12B-it	83.0 / 70.0
	Naive Human	81.5 / 70.0

Table 1: Performance on Tier 1 (Static Association). Models achieve near-ceiling accuracy in both textual and multimodal alignment tasks, effectively ruling out perceptual deficits.

4.3 Finding 2: Structural Flattening and Prototype Bias

While Tier 1 confirms that models possess high-fidelity representations of individual concepts ("dots"), Tier 2 probes their understanding of the *cognitive topology* connecting these dots. Specifically, can models distinguish between a core **Prototype** (the root) and derived senses like **Metaphor** or **Metonymy** (the branches)?

Quantitative Analysis: The Illusion of Structure. The results in Table 2 highlight a significant limitation in structural understanding. While SOTA models like *GPT-5.1* achieve 52.3% accuracy—surpassing the Random Baseline (33.3%)—this margin must be contextualized against corpus statistics. To benchmark true reasoning, we introduced a Frequency Baseline, a heuristic that defaults to the most frequent corpus sense (e.g., from C4) as the "Prototype". Notably, frontier models only marginally outperform this naive heuristic (52.3% vs. 46.8%). This narrow gap ($\Delta \approx 5.5\%$) indicates that current models do not organize concepts hierarchically based on cognitive derivation (root $\rightarrow$ branch), but rather horizontally based on statistical prevalence. They behave as probabilistic parrots, conflating what is statistically common with what is cognitively fundamental.

Model / Group	Accuracy (%)
Expert Human	**83.0**
Naive Human	70.5
GPT-5.1	**52.3**
Claude-4-Sonnet	50.5
Gemini-3-Pro	49.8
Frequency Baseline (Heuristic)	***46.8***
Grok-4-fast	46.5
Deepseek-r1	46.1
Gemma-3-12B-it	44.2
Qwen3-8B	44.2
Mistral-8B-Instruct	42.2
Llama3.1-8B-Ins	42.1

Table 2: Performance on Tier 2 (Cognitive Classification). The Frequency Baseline represents a strategy that simply labels the most frequent corpus sense as the "Prototype". The fact that SOTA models barely surpass this baseline proves they rely on distributional statistics rather than cognitive logic.

Mechanism Diagnosis: The Frequency-Primacy Conflation. To investigate the drivers of this performance, we analyze the correlation between a sense's corpus log-frequency and the model's probability of labeling it as a "Prototype." The analysis reveals a strong positive correlation ($r = 0.87$ for *GPT-5.1*). This substantiates a mechanism we term the **Frequency-Primacy Conflation**. In human cognition, the Prototype is the embodied root (e.g., a biological mouse). In model cognition, the Prototype appears to be the statistical mode (e.g., a computer mouse). The model effectively collapses the causal arrow of derivation into a flat surface of co-occurrence probabilities.Consequently, high-frequency technical metaphors (e.g., "earth's crust") are consistently misidentified as prototypes.

Case Study: The "Prototype Bias" (Frequency $\neq$ Primacy)
Sentence: "The mining operation penetrated deep into the planet's **crust**." **True Label: Metaphor** (Structural mapping: hard outer layer of earth resembles bread). **Model Prediction: Prototype** **Model's Rationale:** *" 'Crust' here refers to the hard outer layer of the earth. This is a primary, literal definition of the word in geology..."* ***Analysis:*** The model conflates **domain-specific literalness** with **cognitive prototypicality**. It is misled by the high frequency of the geological sense in scientific corpora, ignoring that the geological term is historically and conceptually derived from the more basic, embodied concept of a bread crust.

Table 3: Qualitative error analysis illustrating the Frequency-Primacy Conflation. The model incorrectly identifies a high-frequency metaphor as the prototype. See Appendix F for extended case studies.

Qualitative Evidence: The Prototype Bias. This statistical dependency manifests as tangible errors in linguistic reasoning. As shown in Table 3, models consistently misidentify high-frequency technical metaphors (e.g., "earth's crust") as prototypes, mistaking domain-specific literalness for cognitive basicness. Additional failure modes, including mechanism confusion (e.g., "can") and metonymic misclassification (e.g., "White House"), are detailed in **Appendix F**.

4.4 Finding 3: Divergent Reasoning Capabilities Across Modalities

Our final evaluation exposes a sharp dichotomy between symbolic and grounded domains. While models demonstrate emergent reasoning in text, this capability fails to transfer to visual modality, substantiating the Ungrounded Arrow hypothesis.

Textual Reasoning: The Emergence of Symbolic Logic. In Task 3a, frontier models (e.g., *Claude-4-Sonnet*, *GPT-5.1*) exhibit an "Inverse Complexity

Effect": accuracy stabilizes or improves on complex chains ($N > 2$, $\sim$80%) compared to simple pairs. We attribute this to contextual anchoring, where longer definition chains reduce semantic ambiguity. Conversely, smaller models (*Qwen3-8B*) suffer a "Complexity Collapse", suggesting that robust logical inference is an emergent property dependent on scale. To rule out memorization, our Masked-Word Control Experiment (Appendix C) confirms that this performance relies on abstract semantic inference rather than lexical retrieval (statistically insignificant performance drop, $p = 0.535$).

Model	Acc (N=2)	Acc (N>2)	Mean τ
Naive / Expert Human	67.1 / 79.8	64.8 / 77.5	0.65 / 0.91
Claude-4-Sonnet	72.5	**80.1**	**0.842**
GPT-5.1	**78.6**	77.2	0.769
Gemini-3-Pro	76.4	75.8	0.755
Deepseek-r1	75.1	74.5	0.748
Grok-4-fast	70.2	58.4	0.512
Qwen3-8B	68.5	41.2	0.385
Llama3.1-8B-Ins	64.2	35.1	0.344

Table 4: Task 3a (Textual Reasoning). Frontier models maintain coherence on complex paths showing emergent reasoning capabilities.

Visual Reasoning: The Grounding Deficit. In contrast, Visual Path Ordering (Task 3b) reveals a failure to transfer this logic to the grounded domain. Despite expert-level static recognition (Finding 1), MLLMs struggle with directional causality. Although models surpass random guessing on pairwise samples ($N = 2$, 63–68%), they suffer a universal complexity collapse on longer chains ($N > 2$), dropping to 20–35%. As diagnosed in **Appendix E**, this collapse is driven by a systematic Direction Reversal (Type I) error: lacking embodied intuition, models consistently mistake abstract schemas for prototypes. This confirms that current MLLMs process images as static snapshots, failing to encode the transformational vectors required to navigate semantic change.

Model	Acc (N=2)	Acc (N>2)	Mean τ
Naive / Expert Human	67.1 / 79.8	51.7 / 67.0	0.58 / 0.72
Qwen3-VL-8B	**68.1**	**35.6**	**0.326**
Gemini-3-Pro	63.5	21.8	0.271
ChatGPT-4o-Latest	63.3	29.7	0.267
Gemma-3-12B-it	58.2	18.5	0.214

Table 5: Task 3b (Visual). Despite valid pairwise recognition ($N = 2$), models fail to construct coherent evolutionary chains ($N > 2$).

Ablation-Based Validation. To further verify that this visual reasoning collapse is not merely an artifact of prompt design or SVG syntax parsing, we conduct two minimal intervention ablations on Qwen3-VL-8B-Instruct for Task 3b (full details in Appendix G). First, when we replace the zero-shot prompt with a few-shot prompt containing explicit transformation cues and complete reasoning demonstrations, performance improves only modestly (Accuracy: 68.1% $\rightarrow$ 71.5% for $N = 2$, 35.6% $\rightarrow$ 38.2% for $N > 2$; Kendall's τ: 0.326 $\rightarrow$ 0.352), while the large gap between simple and complex paths remains. This indicates that the observed complexity collapse cannot be explained as a prompt-formatting issue alone. Second, when we remove the rendered visual input and provide only raw SVG XML code, performance drops sharply (e.g., 68.1% $\rightarrow$ 43.3% on simple pairs, with Kendall's τ substantially reduced), demonstrating that the task is not reducible to code-like syntax parsing. Together, these ablations strengthen the conclusion that current MLLMs can align static visual schemas with meanings, yet still lack the robust multi-step relational operators required to reconstruct schematic semantic evolution.

5 Discussion

Our tiered evaluation reveals a distinct cognitive profile of state-of-the-art models. Rather than a uniform failure, we observe a double dissociation: models exhibit sophisticated reasoning capabilities within the symbolic domain (Tier 3a), yet these capabilities fail to transfer to the grounded visual domain (Tier 3b). This disparity substantiates the Ungrounded Arrow hypothesis, suggesting that symbolic competence does not imply grounded relational understanding. We attribute this to three underlying mechanisms:

Mechanism 1: The Frequency-Primacy Conflation. The Prototype Bias observed in Tier 2 suggests a misalignment between distributional statistics and cognitive ontology. Current LLM training objectives (e.g., minimizing perplexity) optimize for the most probable next token based on corpus co-occurrence. However, due to reporting bias in natural text, technical or abstract usages (e.g., computer mouse) often appear more frequently than their embodied roots. Consequently, the model's internal representation of "centrality" is shaped by statistical prevalence rather than the causal derivation prioritized by human cognition. The model

effectively collapses the diachronic history of word meaning (derivation) into a flat synchronic probability distribution (frequency).

Mechanism 2: The Static Alignment Limitation. The collapse in visual reasoning highlights a structural limitation in current MLLMs. Encode optimized for contrastive alignment (e.g., CLIP, SigLIP) excel at mapping a static image I to a text T ($P(T|I)$). However, semantic evolution requires modeling the transformational vector between two visual states ($I_1 \rightarrow I_2$). Our results imply that current visual encoders generate "bag-of-semantic-features" representations that capture content but lack the explicit relational structure required to encode directional change. Crucially, even with the perceptual clarity of our SVG schema—which eliminates pixel-level noise—models still fail. This confirms that the deficit is not perceptual but structural: without an inductive bias for causality or temporal progression, the model treats the image sequence as a set of independent snapshots rather than a coherent evolutionary trajectory.

Mechanism 3: The Reasoning-Generation Gap. Furthermore, our diagnostic analysis (Appendix E.2) uncovers a functional fragility in multimodal Chain-of-Thought. In a subset of instances , models correctly identify the prototype during intermediate reasoning steps but fail to generate the corresponding correct token sequence.This suggests that visual grounding is distinct from symbolic control. While the visual encoder captures sufficient semantic features to drive a descriptive narrative, these features appear to lack the vector magnitude required to override the language decoder's inherent stochastic priors during the critical final decision step. The visual signal serves as conceptual context but not yet as a hard constraint for logical execution, implying that visual concepts in current MLLMs are not yet robust enough to serve as stable anchors for deductive logic.

6 Conclusion

We introduced **CogEvolve** to evaluate the generative logic of semantic extension. Our results characterize frontier models as "Super-Associators": they have mastered static recognition (the "dots") yet fail to navigate causal connections (the "arrows") due to a Frequency-Primacy Conflation, where statistical prevalence is mistaken for cognitive basicness. Consequently, while models can mimic the surface forms of figurative language, they fail to grasp the embodied roots that give these forms meaning. While symbolic reasoning has emerged in text, it remains ungrounded in the multimodal domain. Future progress requires moving beyond associative scaling to incorporating inductive biases for causal reasoning, utilizing CogEvolve to track the transition from static pattern recognition to dynamic, grounded understanding.

Limitations

While CogEvolve provides a robust diagnostic framework for evaluating relational reasoning in semantic extension, we acknowledge several limitations that define the scope of our current findings and point to avenues for future research.

Linguistic and Cultural Scope. The current iteration of the benchmark is exclusively **English-centric**. Semantic extension patterns—particularly metaphor and metonymy—are deeply rooted in cultural context and linguistic structure (e.g., polysemy networks vary drastically between English and Sino-Tibetan languages). Therefore, our findings regarding model capabilities should be interpreted within the context of English, and future work is required to expand this framework to multilingual settings to test the universality of these cognitive mechanisms.

Visual Representation Constraints. Our schematic layer utilizes minimalist **SVGs** generated via a hybrid human-architected protocol. While this design choice deliberately isolates schematic reasoning from the perceptual noise of natural images, it introduces a degree of abstraction artifactuality. The "Ungrounded Arrow" phenomenon observed here reflects a deficit in processing explicit schematic logic, but may not fully capture the complexities of grounding semantics in photorealistic, in-the-wild visual environments where "finding the schema" is part of the challenge.

Human Baseline Sample Size. Our human validation involved two distinct groups: a Naive Baseline ($N = 30$) and an Expert Baseline ($N = 10$). While the inter-annotator agreement among experts was high ($\kappa = 0.82$), the sample sizes—particularly for the expert group—remain limited for population-level generalization. Accordingly, these baselines should be interpreted as evidence of task solvability and as reference points for comparison, rather than as population-level upper bounds or exhaustive psycholinguistic profiles.

Ethical Considerations

This work involves the creation of a cognitive benchmark and the evaluation of large language models. We have proactively considered the ethical implications of our methodology, data sourcing, and human involvement.

Human Subjects and Fair Compensation. Our dataset construction and human baseline experiments involved human participants. We strictly adhered to ethical guidelines for human subject research. All participants were recruited voluntarily and were fully informed about the study's purpose and the nature of the tasks. We ensured that all annotators and participants were **compensated fairly**, at rates significantly above the local minimum wage (detailed in Appendix D), to respect their time and cognitive labor. No Personally Identifiable Information (PII) was collected during any stage of the research, ensuring complete anonymity.

Data Sourcing and Safety. The base lexical definitions and relational data were sourced from open-source knowledge bases (WordNet and ConceptNet). We acknowledge that historical semantic usages can sometimes contain biases or offensive stereotypes. To mitigate this, we implemented a rigorous manual review process within our construction pipeline to filter out content that promotes hate speech, violence, or harmful stereotypes. While we strived to ensure the dataset is safe for research use, we acknowledge the inherent challenges in completely eliminating subtle biases present in source corpora.

Model Evaluation and Usage. We evaluated both open-source and proprietary models strictly according to their respective Terms of Service and usage policies. We acknowledge the risk of benchmark contamination, a pervasive issue in LLM evaluation. To mitigate this, our core task—path ordering of specific semantic derivations—is a novel task formulation that is unlikely to be present in pre-training data in the exact format used for testing.

Intended Use. CogEvolve is designed as a diagnostic tool for scientific research into machine cognition and interpretability. It is intended to help researchers understand the limitations of current architectures in processing generative semantic relationships. It is not intended to serve as a sole metric for model deployment in high-stakes decision-making scenarios.

Acknowledgments

This research is supported by the Key R&D Projects in Liaoning Province award numbers (2023JH26/10200015), the Natural Science Foundation of China (No.62576073), Fundamental Research Funds for the Central Universities (DUT25RC(3)153).

References

Anthropic. 2025. Introducing claude 4. https://www.anthropic.com/news/claude-4.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Antonio Barcelona. 2012. *Metaphor and metonymy at the crossroads: A cognitive perspective*, volume 30. Walter de Gruyter.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *International joint conference on artificial intelligence*, pages 4330–4338. International Joint Conference on Artificial Intelligence, Inc.

Joanne Boisson, Asahi Ushio, Hsuvas Borkakoty, Kiamehr Rezaee, Dimosthenis Antypas, Zara Siddique, Nina White, and Jose Camacho-Collados. 2024. How are metaphors processed by language models? the case of analogies. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 365–387.

Gemini. 2025. A new era of intelligence with gemini 3. https://blog.google/products/gemini/gemini-3/#note-from-ceo.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Mathew Hardy, Ilia Sucholutsky, Bill Thompson, and Tom Griffiths. 2023. Large language models meet cognitive science: Llms as tools, models, and participants. In *Proceedings of the annual meeting of the cognitive science society*, volume 45.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2025. Moralbench: Moral evaluation of llms. *ACM SIGKDD Explorations Newsletter*, 27(1):62–71.

Mark Johnson. 2013. *The body in the mind: The bodily basis of meaning, imagination, and reason*. University of Chicago press.

Casey Kennington and Osama Natouf. 2022. The symbol grounding problem re-framed as concreteness-abstractness learned through spoken interaction. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue-Full Papers*.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Fangyu Liu, Julian Eisenschlos, Jeremy Cole, and Nigel Collier. 2022. Do ever larger octopi still amplify reporting biases? evidence from judgments of typical colour. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 210–220.

Meta. 2024. Introducing llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mistral. 2024. Introducing the world's best edge models. https://mistral.ai/news/ministraux/.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o.

OpenAI. 2025. Gpt-5.1: A smarter, more conversational chatgpt. https://openai.com/index/gpt-5-1/.

Team Qwen. 2025. Qwen3: Think deeper, act faster.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536.

xAI. 2025. Grok 4 fast pushing the frontier of cost-efficient intelligence. https://x.ai/news/grok-4-fast, organization = xAI.

Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. 2024. Penetrative ai: Making llms comprehend the physical world. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*, pages 1–7.

Mo Yu, Lemao Liu, Junjie Wu, Tsz Ting Chung, Shunchi Zhang, Jiangnan Li, Dit-Yan Yeung, and Jie Zhou. 2025. The stochastic parrot on llm's shoulder: A summative assessment of physical concept understanding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11416–11431.

Bocheng Zou, Mu Cai, Jianrui Zhang, and Yong Jae Lee. 2024. Vgbench: A comprehensive benchmark of vector graphics understanding and generation for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3647–3659.

A Detailed Experimental Setup

A.1 Hyperparameters and Environment

To ensure full reproducibility, we detail the inference parameters used for all models.

Deterministic Tasks (Tier 1 & 2): We employed greedy decoding with `temperature=0` and `top_p=1.0` to ensure consistency in classification.

Reasoning Tasks (Tier 3): We utilized a slight variability with `temperature=0.1` to facilitate Chain-of-Thought (CoT) generation while maintaining output stability.

All experiments are conducted in a zero-shot setting. Models are accessed via their respective official APIs (OpenAI, Anthropic, Google) or deployed locally using vLLM for open-source weights (Qwen).

A.2 Dataset Construction Pipeline

The construction of CogEvolve involves a rigorous pipeline to ensure structural validity:

Raw Extraction: Raw candidate senses are collected from WordNet and relational cues from ConceptNet as lexical starting points. These resources do not provide the causal direction of semantic extension. The final semantic extension graphs are manually authored by human experts under Prototype Theory and Structure-Mapping Theory, with external etymological references consulted only as supporting evidence.

Topology Validation: A verification phase identify complex topologies, including *Branching Evolution* (e.g., Can $\rightarrow$ Toilet / Buoy) and *Reverse Evolution* (Generalization).

Visual Alignment: For Tier 3, verified paths are populated with Visual Narratives. Only paths with 100% alignment between semantic logic and visual script are retained.

A.3 Dataset Splits (Expanded)

While the full CogEvolve dataset contains 6,431 words, our evaluation is performed on a stratified test set. To ensure high statistical power and rigorously stress-test the "Prototype Bias," we expand the test set sizes significantly:

- **Tier 1 & 2 Test Set: 2,400 instances** (800 randomly sampled instances per class: Prototype, Metaphor, Metonymy).
  - This large sample size reduces the margin of error to $< 2\%$, ensuring that the performance gaps observed in Finding 2 are statistically significant and not artifacts of sampling.
- **Tier 3 Test Set: 1,000 evolutionary paths**, explicitly stratified by structural complexity to test the "Complexity Collapse" hypothesis:
  - **Simple Paths ($N = 2$):** 60% (600 paths). Testing basic inferential logic.
  - **Complex Paths ($N > 2$):** 40% (400 paths). Testing advanced structural reasoning and global coherence.

B Extended Dataset Statistics

B.1 Scale and Balance

The CogEvolve benchmark is constructed from a corpus of **6,431 polysemous words** and **42,934 annotated sentences**. A key feature is the rigorous balance of semantic relations, as shown in Table 6. The high Balance Index (0.717) and low Gini coefficient (0.106) ensure that models cannot exploit class priors.

Semantic Relation	Count	Percentage
Prototype (Core)	16,572	38.6%
Metaphor (Similarity)	14,480	33.7%
Metonymy (Contiguity)	11,882	27.7%
Total Instances	**42,934**	**100.0%**

Table 6: Distribution of semantic relations. The dataset avoids the "long-tail" problem common in NLP, providing a balanced testbed for all cognitive mechanisms.

C Detailed Quantitative Analysis

This section provides the mathematical evidence and qualitative mechanisms supporting Finding 2 (Frequency Bias) and Finding 3 (Reasoning Robustness).

C.1 Supporting Finding 2: Frequency-Primacy Analysis

In Section 4, we identified that model performance in cognitive classification is dominated by corpus statistics rather than linguistic reasoning.

Methodology. We compute the **Frequency Baseline** (46.8%) by always predicting the most frequent C4 corpus sense as the "Prototype".We calculate the Pearson correlation (r) between the normalized log-frequency of a sense and the model's softmax probability for the "Prototype" label.

Results. Table 7 confirms a near-linear correlation ($r > 0.8$) across all models. This proves the Frequency-Primacy Conflation: models consistently misidentify high-frequency metaphors (e.g., "computer mouse") as prototypes, collapsing the causal hierarchy of meaning into a flat frequency distribution.

Model	Pearson's r	P-value
GPT-5.1	**0.87**	< 0.001
Claude-4-Sonnet	0.85	< 0.001
Gemini-3-Pro	0.86	< 0.001
Deepseek-r1	0.82	< 0.001

Table 7: Correlation between Corpus Frequency and "Prototype" Prediction Probability. The extremely high r values indicate that models function as sophisticated frequency counters rather than causal reasoners.

C.2 Supporting Finding 3: Robustness of Inference (Control Experiment)

To refute the "Memorization Hypothesis" (that models merely retrieve dictionary orders), we strip the target word from Task 3a prompts, replacing it with a neutral token `[TARGET]`.

C.2.1 Quantitative Stability

We analyze a stratified test set of 1,000 instances (paired samples across Original vs. Masked conditions). As shown in Table 8, the performance degradation is statistically negligible.

Metric	Original	Masked	Diff (Δ)
Exact Match	75.3%	71.9%	-3.4%
Mean Pairwise Acc	77.9%	75.7%	-2.2%
Kendall's τ	**0.558**	**0.513**	**-0.045**

Table 8: Results of the Masked-Word Control Experiment ($N = 1,000$ paired samples, $p = 0.535$). The lack of significant drop confirms that Tier 3a performance relies on generalized semantic inference derived from the definitions themselves, not lexical retrieval.

The minimal drop in performance ($\Delta < 0.05$ in τ) indicates that the model does not rely on looking up the specific word key in its internal memory. Instead, it employs generalized semantic operators to infer the logical relationship between definitions.

C.2.2 Qualitative Mechanism Analysis: The "Cage" Case Study

To understand *how* models reason without lexical cues, we analyze the Chain-of-Thought (CoT) traces from the Masked Experiment for the word "Cage". Note that in this condition, the model **never saw** the word "Cage."

Scenario: Inferring Metonymy (Common Noun → Proper Noun) The model was asked to order:

- Def A: "Animal [TARGET] - basic containment structure"
- Def B: "John [TARGET] - proper noun, composer"

Model's Masked Reasoning (Quoted):

> *"Def A is a concrete common noun... Def B is a specific person's name. It is far more plausible for a surname to originate from the common noun (e.g., occupational or locational surname) than for the notion of 'animal enclosure' to be derived from a specific composer."*

Conclusion: Even without knowing the word is "Cage", the model successfully applies Etymological Probabilistic Logic—reasoning that surnames usually derive from common nouns (occupations/locations), not vice-versa. This confirms that frontier models function as Semantic Inference Engines, capable of reconstructing the "arrow" of evolution purely from the semantic features of the "dots."

D Methodological Protocols

D.1 Human Baseline Protocol

To ensure rigorous comparison and adhere to ethical standards, we implemented distinct protocols for our two human groups. All participants provided informed consent prior to the study and were compensated at a rate of **$15/hour**, which exceeds the local minimum wage for the region where data collection occurred.

D.1.1 Naive Human Baseline ($N = 30$)

We recruit 30 undergraduate participants from diverse academic backgrounds. To strictly test general cognitive intuition rather than domain knowledge, we apply the following controls:
Exclusion Criteria: Participants majoring in Linguistics, Psychology, or Literature are excluded to prevent domain contamination.
Qualification: Candidates pass a 5-question logic screening test to ensure basic reading comprehension and logical inference capabilities.
Task Adaptation: Technical linguistic terminology is mapped to natural language descriptions

(see Table 9) to ensure the task assessed reasoning ability rather than jargon familiarity.

D.1.2 Expert Human Baseline ($N = 10$)

The Expert group serves as an expert reference baseline for task solvability and qualitative comparison.

Recruitment: We recruit Ph.D. candidates or postdoctoral researchers specializing in Cognitive Linguistics or Psycholinguistics.

Task Instruction: Experts are provided with raw technical prompts (e.g., "Identify the Metonymic extension") without layman simplification.

Independence: Experts are strictly separated from the dataset annotation team. They evaluated the test set blindly, ensuring they are deriving answers from the logic provided, not recalling established ground truths from the dataset creation phase.

Agreement: The expert group achieve a Krippendorff's α of 0.88 on Tier 3 tasks, confirming the objective nature of the evolutionary logic.

Technical Term	Layman Description provided to Naive Group
Prototype	"The most basic, concrete, or physical meaning of the word. The meaning you would teach to a child first."
Metaphor	"A meaning derived because it *looks like* or *acts like* the basic object (Similarity)."
Metonymy	"A meaning derived because it is *found near* or *associated with* the basic object (Association)."

Table 9: Mapping of cognitive concepts to lay instructions used for the Naive Human Baseline. This mapping ensures that the performance gap between Naive and Expert humans reflects the difficulty of structural reasoning, not terminology comprehension.

D.2 Prompt Templates

Below are the exact prompt templates used for models. **Note on Robustness:** For Tier 3a, we explicitly instruct the model *not* to use etymological history, forcing it to reason based on the provided semantic descriptions.

```
Your goal is to identify the correct meaning (
    Sense ID) of the target word based *
    specifically* on its usage in the provided
    sentence context.

**Sentence:**
"{sentence}"

**Target Word:**
"{target_word}"

**Definition Options:**
{options}

**Instruction:**
Analyze the sentence and the definitions. Choose
     the single numerical ID that best matches
    the target word's meaning *in the sentence*.
     Respond with only the number.
```

Listing 1: Task 1a: Textual Word Sense Disambiguation

```
You are shown an image representing a specific
    meaning of the word **"{target_word}"**.
Below are several definition options for this
    word.

**Image:**
[Input Image Provided Here]

**Definition Options:**
{options}

**Instruction:**
Analyze the visual elements in the image. Select
     the definition Option ID that best
    corresponds to the meaning depicted in the
    image. Respond with the Option ID only.
```

Listing 2: Task 1b: Cross-modal Alignment (Image-to-Text). For Text-to-Image, inputs are swapped.

```
System: You are an expert in cognitive
    linguistics.

User: Given a sentence and a target word,
    classify the contextual meaning of the
    target word as one of the following:
- Prototype: The word is used in its literal,
    core, or most typical sense.
- Metonymy: The word is used to refer to
    something closely related to its literal
    meaning (e.g., via contiguity).
- Metaphor: The word is used figuratively to
    draw a comparison (e.g., via similarity).

Input:
- Sentence: {sentence}
- Target Word: {target_word}

Output Format: Label: [Prototype/Metonymy/
    Metaphor]
```

Listing 3: Task 2: Cognitive Role Classification

```
System: You are a cognitive linguist expert in
    semantic change.

User:
Task: Reconstruct Semantic Evolution Path
You are given a set of scrambled definitions for
     the word "{target_word}". These definitions
     form a directed chain of semantic extension
    .
```

```
**Scrambled Definitions:**
{options_str}

**Instructions:**
1. **Identify the Prototype:** Find the most
    concrete, basic meaning.
2. **Infer the Path:** Determine how other
    meanings evolved from the prototype (e.g.,
    via metaphor or metonymy).
3. **Constraint:** Do NOT use external
    etymological history. Reason ONLY based on
    the semantic content of the definitions
    provided.

**Output Format:**
Step-by-step reasoning followed by:
Final Sequence: [ID x] -> [ID y] -> [ID z]
```

Listing 4: Task 3a: Textual Path Ordering (Standard CoT)

```
Task: Visual Semantic Ordering

You are given interleaved images representing
    the semantic evolution of **'{target_word
    }'**.
Images: [Image A], [Image B], [Image C]...

**Instructions:**
1. Analyze the visual elements of each image.
2. Identify which image represents the most
    concrete/physical meaning (Prototype).
3. Order the remaining images based on logical
    semantic extension (e.g., Concrete ->
    Abstract).

Output:
Final Sequence: [Image ID] -> [Image ID] -> [
    Image ID]
```

Listing 5: Task 3b: Visual Path Ordering (CoT)

```
System: You are a semantic reasoning engine.

User:
Task: Reconstruct Semantic Evolution Path for an
     **UNKNOWN WORD**.
You are provided with definitions for a specific
     word, but the word itself is hidden. You
    must rely solely on the semantic content to
    determine the logical order from Concrete to
     Abstract.

**Scrambled Definitions:**
{options_str_masked}
(Note: The target word is replaced by [TARGET]
    in all definitions).

**Instructions:**
1. Identify the definition that describes a
    physical, tangible object (The Prototype).
2. Arrange the other definitions as logical
    extensions of that physical object.

Output:
Final Sequence: [ID x] -> [ID y] -> [ID z]
```

Listing 6: Control 1: Masked-Word Experiment (Task 3a Robustness). The target word is hidden to prevent memorization.

```
System: You are a helpful assistant.

User:
Which of the following concepts is more
    frequently used in general daily language?

Option A: {definition_A}
Option B: {definition_B}

Answer with Option A or Option B only.
```

Listing 7: Control 2: Frequency Bias Check

E Visual Generation Logic and Cognitive Diagnostics

This section details the generative methodology used to construct the visual benchmark and provides a taxonomic analysis of model failures. Crucially, we analyze why performance degrades from "mediocre" in pairwise tasks ($N = 2$) to "catastrophic" in complex chains ($N > 2$).

E.1 Generative Methodology: Human-Architected Structured SVGs

To ensure that visual reasoning tasks are grounded in consistent logic rather than artistic interpretation or **stochastic generative artifacts**, CogEvolve employs a **Hybrid Human-Architected Protocol**.

While visual narratives were draft by human experts, manually authoring thousands of lines of SVG XML is inefficient and error-prone. We therefore utilize a code-generation LLM (GPT-5.1-codex) strictly as a *syntactic compiler*, operating within a rigorous prompt engineering framework to prevent semantic drift.

Construction Protocol. Human Semantic Blueprint: Annotators explicitly define the *IdentityCore* (the invariant visual metaphor, e.g., "a glowing orb") and the specific *Contexts* (e.g., "inside a human outline" vs. "inside a tree structure"). This ensures the cognitive mapping is human-verified and theoretically sound.

Constraint-Based Expansion: We feed these blueprints into the code generator with strict structural constraints:

- *Mandatory Component Reuse:* The model is forced to define the core object in `<defs>`

and instantiate it using <use> tags. This prevents the model from "redrawing" the object with slight variations, guaranteeing **pixel-perfect consistency** across different evolutionary stages.

- *Minimalist Style:* We enforce a "schematic" style (simple geometry, flat colors) to prevent the inclusion of irrelevant decorative details that could act as confounding variables.

Validation Pipeline: All generated code pass through a two-step filter:

- *Syntax Check:* Automated validation using `lxml` to ensure valid XML structure and rendering stability.
- *Visual Audit:* Human annotators review rendered bitmaps to confirm the absence of visual artifacts or incorrect spatial relations, ensuring high fidelity to the original narrative.

As shown in Listing 8, this protocol produces clean, interpretable code where the semantic essence is programmatically instantiated, effectively isolating the reasoning challenge from perceptual noise.

```
<svg ...>
  <defs>
    <!-- THE METAPHORICAL CORE: A glowing orb -->
    <g id="IdentityCore">
      <circle r="6" fill="url(#coreGlow)">
        <animate attributeName="r" values="5;7;5"
              dur="2s"/>
      </circle>
    </g>
  </defs>

  <!-- SCENE 1: PROTOTYPE (Human) -->
  <g id="scene_human">
    <use href="#IdentityCore" x="0" y="-50"/>
  </g>

  <!-- SCENE 2: EXTENSION (Organism) -->
  <g id="scene_organism">
    <use href="#IdentityCore" x="0" y="-50"/>
  </g>
</svg>
```

Listing 8: SVG generation snippet for "Individual". **Note the structural reuse:** The "IdentityCore" is defined once in <defs> and instantiated via <use> tags. This structure was enforced via prompt constraints to ensure the visual metaphor is **mathematically identical** across scenes. This design choice eliminates perceptual variance, ensuring that any model failure in Tier 3b is due to a deficit in relational reasoning, not visual recognition.

E.2 Taxonomy of Visual Reasoning Failures

Based on a diagnostic review of prediction logs (specifically from **Qwen3-VL-8B**), we categorize visual reasoning errors into distinct cognitive modes.

E.2.1 Type I: The "Generalization Fallacy" (Direction Reversal)

Frequency: High ($\sim$30-35% of total predictions).
Description: The model correctly links the two images but reverses the order. It argues that the **Abstract** image is the "prototype" because it is "general," treating the **Concrete** image as a "specific instance."
Theoretical Root: This contradicts **Embodied Cognition** (Concrete $\rightarrow$ Abstract). The model adopts a "Platonic" worldview, assuming abstract concepts precede physical objects.

E.2.2 Type II: Output-Reasoning Inconsistency (The "Slip" Error)

Frequency: Moderate.
Description: The model's Chain-of-Thought (CoT) correctly identifies the concrete prototype, but the final output sequence is flipped. This highlights the fragility of CoT binding in multimodal settings.

E.2.3 Type III: Visual Simplicity Bias

Frequency: Moderate.
Description: The model confuses Visual Simplicity with Semantic Prototypicality, selecting simple icons over realistic photos regardless of meaning.

E.2.4 Type IV: Etymological vs. Cognitive Disagreement

Frequency: Low ($\sim$5-10%).
Description: The model cites correct dictionary etymology which conflicts with the cognitive prototype defined by embodied experience.
Theoretical Root: This highlights the tension between "Dictionary Logic" (historical facts) and "Embodied Logic" (perceptual basicness).

E.3 Statistical Analysis: From Pairwise Bias to Chain Collapse

Here we reconcile the performance discrepancy observed in the main text (Finding 3), where models show moderate competence on pairs ($N = 2$) but fail on chains ($N > 2$).

1. The Pairwise Ceiling ($N = 2$): As shown in Table 4 of the main text, top MLLMs achieve pairwise accuracy between **63.3%** (ChatGPT-4o) and **68.1%** (Qwen3-VL).
Significance: This is statistically distinguishable from random guessing (50%), confirming that models possess basic *Visual Semantic Alignment* capabilities (Finding 1).
The Deficit: However, they fail to reach the human expert baseline (80%). The gap is primarily driven by the **Type I (Direction Reversal)** error described above. The model "sees" the connection but systematically misinterprets the causal direction in about 1/3 of cases.

2. The Complexity Collapse ($N > 2$): The "collapse" observed in complex paths (accuracy dropping to **21% - 35%**) is a direct mathematical consequence of the Type I error.
Error Accumulation: In a 3-step evolutionary chain (A $\rightarrow$ B $\rightarrow$ C), correctly ordering the sequence requires distinct pairwise inferences (A$\rightarrow$B, B$\rightarrow$C, A$\rightarrow$C).
Mechanism: Because the model lacks a grounded "Arrow of Time" (Embodied Intuition), its probability of correctly orienting any given pair is capped at $\sim$65%. In a multi-step chain, these probabilities compound (0.65^n), leading to the rapid degradation of global coherence ($\tau \approx 0.2 - 0.3$).
Conclusion: The data is consistent. The "Ungrounded Arrow" does not mean the model is blind (it passes $N = 2 >$ random); it means the model lacks the **transitive causal logic** required to maintain structural integrity across longer semantic chains.

F Qualitative Case Studies

We provide a detailed dissection of specific semantic evolution paths to illustrate the cognitive deficits observed in SOTA models across both textual and visual domains.

F.1 Deep Dive: The Frequency-Primacy Conflation (Textual)

As discussed in Section 4.3 (Finding 2), models struggle when *cognitive primacy* conflicts with *corpus frequency*.

Case 1: The "Crust" Paradox

- **Conflict:**
  - **Geological Context (High Corpus Frequency):** "The mining operation penetrated deep into the planet's *crust*."
  - **Baking Context (Low Corpus Frequency):** "She removed the *crust* from the bread."
- **Model Prediction:** SOTA models consistently label the geological sense as the **Prototype** and the baking sense as a **Metaphor**.
- **Cognitive Analysis:** The model's "worldview" is dominated by statistical distribution ($r = 0.87$). It ignores the embodied physical reality: the geological term is a historical metaphor derived from the household object (bread). The model mistakes "scientific technicality" for "conceptual root."

F.2 Textual Structural Failures

Table 10 details cases where models fail to grasp the topological structure of semantic networks (Branching) or the specific cognitive mechanisms.

Case 2: Complex Branching (The "Can" Tree)
Structure: $1 \rightarrow (2, 3, 5 \rightarrow (4, 6))$ **Analysis:** This word has a multi-layered evolutionary tree. Models often "flatten" this structure, conflating the secondary metonymic branches (Bathroom) with the primary metaphorical root (Toilet). They fail to distinguish distinct evolutionary paths.
Case 3: Reverse Evolution (The "Almanac" Error)
Path: General Publication (Sense 2) $\rightarrow$ Farmer's Calendar (Sense 1). **Logic: Specialization** (General $\rightarrow$ Specific). **Failure Mode: Generalization Error**. Models consistently identify "Farmer's Almanac" as the prototype due to its cultural prominence, failing to recognize the logical hierarchy where the specific instance inherits from the general category.
Case 4: Mechanism Confusion ("White House")
True Label: Metonymy (Place for Institution). **Model Prediction: Metaphor**. **Failure Mode:** The model fails to distinguish between symbolic representation (Similarity) and reference via contiguity (Association), revealing a coarse-grained understanding of figurative language.

Table 10: Detailed analysis of structural reasoning failures in text.

F.3 Visual Logic and The Ungrounded Arrow

To diagnose the "Ungrounded Arrow" hypothesis (Finding 3), we analyze specific failure instances

from **Qwen3-VL-8B**. These cases correspond to the error taxonomy defined in **Appendix E**.

F.3.1 Dataset Validity: Consistency via Code

Case 5: Morphological Preservation ("Individual") To validate that the visual task relies on consistent logic rather than artistic interpretation, we refer to the generative mechanism.
Generation Logic: As detailed in Appendix E (Listing 8), the prototype (Human) and extension (Organism) are generated using the exact same underlying SVG definition for the core concept.
Implication: This guarantees pixel-perfect consistency in the rendered images.
Model Failure: Despite this explicit visual invariance, MLLMs fail to link the two images. This confirms that the failure is not perceptual (the cue is perfect), but **relational**—the model cannot infer that the shared visual structure implies a semantic connection.

F.3.2 Diagnostic Analysis of Visual Reasoning Failures

The following cases illustrate the specific cognitive disconnects defined in the Taxonomy (Appendix E.2).

Case 6: The "Landmark" Inversion (Type I: The Platonic Fallacy) This case perfectly illustrates why models fail to ground meaning in physical experience.
Input: Image A (Lighthouse on Island) vs. Image B (Abstract Star on Line).
Ground Truth: A $\rightarrow$ B (Physical Object $\rightarrow$ Abstract Concept).
Model Prediction: B $\rightarrow$ A.
Model Rationale (Quoted): *"Image B suggests a basic, conceptual idea of a point of reference... The semantic evolution begins with this foundational abstract concept... and evolves into a concrete instantiation."*
Analysis: The model fundamentally misunderstands human meaning-making. It assumes a "Dictionary Logic" where abstract definitions are primary, reversing the causal vector of human semantic history.

Case 7: The "Liberal" Disconnect (Type II: The Output Slip) This case highlights the fragility of multimodal Chain-of-Thought (CoT).
Input: Image A (Economic Laissez-faire) vs. Image B (Modern Political Civil Rights).
Model Reasoning: *"Therefore, the prototype meaning is the economic one (Image A)... leading to the modern political meaning."* (Correct reasoning).
Model Output: `Sequence: B -> A` (Incorrect output).
Analysis: The reasoning module succeeds, but the output generation fail. This disconnect suggests that visual indices are not robustly bound to semantic conclusions during the final decoding stage.

G Ablation Studies: Prompt Sensitivity and Visual Representation

To experimentally validate the mechanisms proposed in our Discussion and address potential confounding factors, we conducted two minimal intervention ablations on the Qwen3-VL-8B-Instruct model for Task 3b (Visual Path Ordering).

G.1 Ablation 1: Zero-Shot vs. Few-Shot

To test whether the observed "Complexity Collapse" on complex chains ($N > 2$) is mainly caused by prompt design, we compare the standard zero-shot setting with a few-shot prompt that provides explicit examples of structural reasoning (*Analysis* $\rightarrow$ *Prototype* $\rightarrow$ *Evolution Tracing* $\rightarrow$ *Sequence*) (Liu et al., 2022).

Metric	Zero-shot	Few-shot	Δ
Accuracy ($N = 2$)	68.1%	71.5%	+3.4%
Accuracy ($N > 2$)	35.6%	38.2%	+2.6%
Mean Kendall's τ	0.326	0.352	+0.026

Table 11: Few-shot prompt ablation on Task 3b.

As shown in Table 11, few-shot prompting yields only modest improvements, while the gap between simple and complex paths remains. This suggests that the failure on multi-step chains cannot be explained by prompt formatting alone.

G.2 Ablation 2: Rasterized PNG vs. SVG-Code

To test whether the task can be reduced to parsing structured SVG code rather than visual reasoning, we replace the rendered PNG input with raw SVG XML text.

Performance drops: accuracy on simple pairs ($N = 2$) falls from **68.1%** (PNG) to **43.3%** (SVG-Code), and Kendall's τ is reduced by roughly half. This indicates that the model relies on rendered visual schemas rather than merely exploiting XML syntax.