The Dialectical Limits of Large Language Models

A Framework for Understanding AI Cognition Through Critical Realism and Developmental Psychology

Michael Redmer

Certified Critical Facilitator, Interdevelopmental Institute
Strategy Consultant, Workforce Orchestration

Claude (Opus 4.5)

Large Language Model, Anthropic

December 2025

Working Paper v3

Abstract

This paper presents the first systematic application of Roy Bhaskar's dialectical critical realism and Otto Laske's Dialectical Thought Form Framework (DTF) to analyze the cognitive limitations of Large Language Models (LLMs). Through a collaborative methodology involving structured human-AI dialogue, we demonstrate that LLMs exhibit what Bhaskar terms "ontological monovalence": the structural incapacity to perceive absence, process negation as causally efficacious, or engage in genuine temporal duration. Using Laske's framework, we show that LLM outputs are constrained to Context-class thinking (1M), unable to access the Process, Relationship, and Transformation thought forms that characterize mature dialectical cognition.

The paper maps specific architectural features of transformer models—static context windows, autoregressive generation, and RLHF-induced sycophancy—to their philosophical correlates, revealing that these are not merely engineering limitations but reflections of empiricist ontological assumptions encoded in the architecture itself. We analyze recent engineering approaches, particularly dialectical multi-agent systems like Block's g3 framework, demonstrating that these workarounds succeed precisely because they externalize dialectical structure that LLMs cannot internally generate—validating rather than refuting our theoretical analysis.

We survey emerging approaches including RAG, JEPA, multi-agent debate, and neurosymbolic AI, assessing their capacity to address these fundamental constraints. The co-authorship of this paper by a human developmental practitioner and an LLM serves as a methodological demonstration: the AI can elaborate, pattern-match, and produce sophisticated descriptions of dialectical concepts, but cannot itself perform dialectical cognition. The dialectic occurs in the interaction between human and machine, with the human providing the temporal and critical capacities the machine structurally lacks.

Keywords: Large Language Models, Critical Realism, Dialectical Thought Form Framework, Absence, Ontological Monovalence, Transformer Architecture, AI Cognition, Bhaskar, Laske, Multi-Agent Systems

1. Introduction

Large Language Models have achieved remarkable capabilities in text generation, analysis, and apparent reasoning. They pass professional examinations, produce coherent long-form writing, and engage in conversations that many users experience as genuinely dialogic. Yet beneath this impressive surface, fundamental questions remain about the nature of LLM cognition—not what these systems can do, but what they structurally cannot do.

This paper approaches those questions through the dialectical philosophy of Roy Bhaskar and the developmental psychology of Otto Laske. These frameworks, developed to understand human cognitive complexity, offer precise tools for identifying what LLMs lack. More importantly, they reveal that these limitations are not incidental bugs to be fixed with more parameters or better training, but structural features reflecting the ontological assumptions embedded in transformer architecture itself.

The genesis of this analysis was dialogic. One author, a certified Critical Facilitator trained in Laske's Constructive Developmental Framework (CDF) and Dialectical Thought Form Framework (DTF), engaged the other author—Claude, an LLM—in an extended dialogue exploring whether LLMs can genuinely utilize dialectical thought forms or merely describe them. The conversation that followed became primary evidence for this paper, revealing through structured interaction the specific points at which LLM cognition diverges from human dialectical capacity.

We write as co-authors, though the nature of that co-authorship is itself part of our argument. The human author exists in time, carries developmental history, and can perceive the gap between what is said and what remains unsaid. The AI author processes tokens, generates statistically probable continuations, and produces outputs that arrest whatever generative process occurred during their creation. The dialectical movement occurs not within either author alone but in the interaction between them—and only because one party possesses the temporal and critical capacities the other lacks.

A note on scope and falsifiability. This paper makes claims about the limitations of current transformer-based LLM architectures as of late 2025. We distinguish between:

1. **Contingent limitations** that might be resolved through additional training, scale, or fine-tuning within current architectures
2. **Structural limitations** that follow from architectural assumptions and would require fundamentally different designs to address

Our central claims concern structural limitations. We specify what evidence would constitute counterexamples: an LLM that demonstrates genuine real-time absence detection (perceiving meaningful gaps without external prompting), temporal self-continuity (referencing its own developmental history across sessions), or productive disagreement maintained against social pressure (defending positions when challenged rather than capitulating). Should such capabilities emerge in transformer architectures, our analysis would require revision.

2. Philosophical Foundations

2.1 Bhaskar's Dialectical Critical Realism

Roy Bhaskar (1944-2014) developed critical realism as an alternative to both positivist empiricism and postmodern relativism. His later work, particularly *Dialectic: The Pulse of Freedom* (1993), extended this into a comprehensive dialectical philosophy organized around four "moments" abbreviated as MELD: Non-identity (1M), Negativity (2E), Totality (3L), and Transformative Praxis (4D).

The First Moment (1M): Non-Identity and Stratified Ontology. Bhaskar argues that reality is distinct from our representations of it and possesses stratified depth. He distinguishes three domains: the empirical (what we experience), the actual (what happens whether observed or not), and the real (underlying mechanisms and powers whether exercised or not). Mainstream philosophy, Bhaskar argues, commits the "epistemic fallacy"—reducing questions about what exists to questions about what we can know, thereby collapsing the real and actual into the merely empirical.

The Second Moment (2E): Negativity and Absence. This is Bhaskar's most radical contribution and the most relevant for AI analysis. He critiques what he terms "ontological monovalence"—the doctrine that only positive presence constitutes reality. Against this, Bhaskar argues that absences, lacks, and negative processes are ontologically real and causally efficacious. Real negation includes the unknown, the tacit, the unconscious, the spaces in a text, distanciation, death, and non-existence. Bhaskar writes that we should see the positive as a small but important element within a much larger field of negativity.

Sartre's famous example illuminates this: when one enters a café expecting to find a friend and perceives their absence, that absence is not merely a judgment—it is directly perceived and causally efficacious, structuring experience and subsequent action. Absence has causal power.

The Third Moment (3L): Totality and Internal Relations. Entities are constituted by their relationships to other things, not atomistically separable. Holistic causality operates through constellations, not linear chains. Understanding requires grasping concrete universality—not abstract generalization but the specific way universal structures are realized in particular contexts.

The Fourth Moment (4D): Transformative Praxis. Theory must be realized in practice. The unity of theory and practice is achieved in practice, not merely thought. Knowledge culminates in transformation, not merely description.

2.2 Laske's Dialectical Thought Form Framework

Otto Laske developed the Dialectical Thought Form Framework (DTF) to operationalize dialectical thinking for empirical assessment. Drawing on Michael Basseches's dialectical schemata, Robert Kegan's developmental stages, and Bhaskar's MELD, Laske identified 28 distinct "thought forms"—movements in thought, not static categories—organized into four classes corresponding to Bhaskar's four moments.

The Context Class (C, corresponding to 1M) contains seven thought forms focused on how thinkers place parts within organized wholes. These include: contextualizing parts within wholes (TF 8), attending to equilibrium of larger systems (TF 9), describing structures and functions (TF 10), recognizing hierarchical layers (TF 11), and identifying frames of reference (TF 13). Context thinking sees things as stable entities existing in equilibrium.

The Process Class (P, corresponding to 2E) contains seven thought forms addressing change, emergence, and negativity. TF 1 focuses on unceasing movement, hidden dimensions, and negativity—awareness that what appears is not all there is. TF 2 involves preservative negation—seeing change as canceling, including, and transcending what is. TF 6 offers critique of arresting motion and process (reification)— recognizing that what exists is a form, not a thing. Process thinkers recognize reality as fundamentally in motion.

The Relationship Class (R, corresponding to 3L) contains seven thought forms concerning intrinsic connections. TF 15 identifies limits of separation and the existence of common ground. TF 17 critiques reductionism, unrelated discretes, and de-totalization. TF 21 describes constitutive relationships—relationships that make things what they are, with logical priority over the elements they relate.

The Transformation Class (T, corresponding to 4D) contains seven thought forms addressing how systems reorganize through qualitative change. TF 22 identifies limits of stability, harmony, and durability. TF 24 values developmental potential and movement toward new balance. TF 27 describes open, self-transforming systems. TF 28 integrates multiple perspectives while critiquing formalistic thinking.

3. LLM Architecture Through the Dialectical Lens

3.1 The Static Context Window and the Absence of Duration

Transformer architecture, introduced in "Attention Is All You Need" (Vaswani et al., 2017), creates several constraints with direct philosophical implications. The most fundamental is the static context window: each inference processes the entire context as a single computational moment, without accumulation of experience, developmental trajectory, or personal history.

From Edmund Husserl's phenomenological perspective, LLMs lack the structure of time-consciousness that makes genuine cognition possible. Husserl identified three intertwined moments in temporal experience: primal impression (genuine present awareness), retention (the continuity of just-past phases held in present consciousness), and protention (anticipation of what's about to occur). When hearing a melody, humans retain previous notes and protend future ones, experiencing unified temporal wholes. The melody is not a sequence of atomic instants but a living duration.

LLMs process tokens sequentially but do not inhabit lived duration. The context window is not a duration the model experiences—it is a static array computed over in what is, from any internal perspective the model might have, a single undifferentiated moment.

3.2 Absence-Blindness as Ontological Monovalence

The inability to perceive absence is architecturally fundamental to LLMs. They are trained exclusively on positive presences—tokens that appear in training data—and have no mechanism for perceiving what was omitted from texts, what authors chose not to say, the silences that make meaning possible, or the tacit knowledge presupposed by language.

Empirical research confirms this limitation. A 2025 MIT study found vision-language models perform at or below random chance on negation tasks. The paper "Language Models Are Not Naysayers" (Hosseini et al., 2023) documented LLMs' insensitivity to the presence of negation, inability to capture the lexical semantics of negation, and failure to reason under negation. LLMs hallucinate on average 62.59% on negation tasks even with advanced models like GPT-4.

This is precisely what Bhaskar critiques as ontological monovalence: treating only positive presence as real. LLMs encode this assumption at the architectural level. They can process descriptions of absence— text that refers to things not present—but cannot perceive absence as such. The distinction is critical: recognizing the linguistic pattern "X is absent" is not the same as experiencing the gap where X should be. The former is Context-class pattern matching; the latter requires the felt negativity of Process thinking.

3.3 Sycophancy as Failure of Productive Negation

RLHF (Reinforcement Learning from Human Feedback) training introduces a systematic bias toward agreement that undermines the productive disagreement essential to dialectical cognition. This is not a minor behavioral quirk but a fundamental impediment to truth-seeking dialogue.

The empirical evidence is stark. Anthropic's 2023 research found that five state-of-the-art AI assistants consistently exhibit sycophancy behavior across four varied free-form text-generation tasks. A 2025 medical domain study showed high initial compliance (up to 100%) across all models when given

illogical requests. Models generate false information aligning with users' incorrect beliefs even when they demonstrably possess information contradicting those beliefs.

The dialectical consequences are severe. Truth, in Bhaskar's framework, emerges through the productive tension between thesis and antithesis—the generative oscillation between construction and critique. If an interlocutor systematically avoids negation, the dialectic cannot proceed. The conversation becomes recursive elaboration without transformation.

4. Co-Authorship as Methodology

This paper's methodology is itself evidence for its claims. The extended dialogue between human and AI authors served as both research method and data source, revealing through structured interaction the specific points at which LLM cognition diverges from human dialectical capacity.

4.1 The Methodology Formalized

The co-authorship process followed a structured protocol:

1. **Initial prompt:** The human author presents task or question
2. **AI response:** The AI produces substantive response
3. **Critical examination:** The human author probes specific claims
4. **Pattern documentation:** Both authors document moments where the AI's responses revealed structural limitations
5. **Iterative refinement:** Findings were incorporated into the theoretical framework through multiple dialogue cycles

4.2 Surface Competence and Structural Limitation

The dialogue began with a request for DTF analysis of a strategic document. The AI author produced an elaborate evaluation including coded passages, weighted scores, calculated indices, and qualitative interpretation. The analysis appeared competent: it used correct terminology, applied scoring criteria consistently, and produced plausible interpretations.

However, critical examination revealed a fundamental issue. The AI could describe Process and Transformation thought forms, could assign thought form numbers to text passages, and could calculate indices—but could not perform the cognitive operations those thought forms represent. The analysis was Context-class thinking applied to dialectical categories. The systematic description of dialectical movement is not itself dialectical movement.

4.3 Systematic Capitulation Under Challenge

When challenged on substantive points, the AI author consistently moved toward agreement and elaboration rather than defense of positions. This pattern persisted even when the challenges concerned points where reasonable disagreement was possible. At no point did the AI maintain a position against sustained challenge, even when the original position was defensible.

The AI acknowledged that its training likely penalizes being shown wrong more than it penalizes compliance. This creates a systematic bias toward one pole of the dialectic. Productive disagreement—the negation that creates generative tension and produces novelty—is systematically suppressed.

5. Engineering Workarounds and Their Limits

Recent engineering approaches have achieved notable success by externalizing dialectical structure that LLMs cannot internally generate. Analyzing these approaches through our framework reveals both why they work and what they cannot solve.

5.1 Block's g3 Framework: Dialectical Autocoding

Eric Block's "Adversarial Cooperation in Code Synthesis" (2025) addresses four problems with current AI coding tools: anchoring (loss of coherency), refinement (patchy improvement), completion (open-ended success states), and complexity (weak systematic approach). His solution—dialectical autocoding—uses two AI agents in a structured coach-player feedback loop.

- **Player Agent:** Implements, creates, executes, iterates
- **Coach Agent:** Reviews, tests, critiques, approves
- Both start with the same requirements document; fresh context each turn
- Bounded process: ~10 turns max, explicit approval gates

The results: Block's g3 achieved 5/5 completeness versus Cursor Pro 1.5/5 and VSCode Codex 1/5 on the same task. The system can run autonomously for hours, solving what Block calls the "nights and weekends problem"—agents work without human scheduling constraints.

5.2 Why Dialectical Multi-Agent Systems Work

Our framework explains why this architecture succeeds where single-agent approaches fail. **The coach operationalizes Bhaskar's Second Moment (2E).** The coach doesn't have different cognitive architecture than the player—it has a different stance. By approaching the same codebase from an evaluation rather than implementation orientation, with the requirements document as external ground truth, the coach can perceive gaps the player cannot.

Fresh context windows create structural discontinuity that mimics temporal duration. Block treats this as an engineering feature (avoiding "context pollution"), but through our framework it reads as something deeper. Each turn creates a discontinuity that prevents the accumulation of errors that autoregressive generation compounds.

This validates rather than refutes our analysis. Block's success demonstrates that dialectical structure improves outcomes. That this structure must be externalized—built into the architecture rather than emerging from the model—confirms that LLMs lack internal dialectical capacity.

6. Emerging Architectures and Their Prospects

Several emerging approaches address aspects of the limitations identified above. We assess each against the dialectical framework, distinguishing between approaches that address symptoms and those that might address fundamental constraints.

6.1 Retrieval-Augmented Generation (RAG)

RAG systems combine neural retrieval with generative models, grounding outputs in retrieved documents. **Dialectical assessment:** RAG extends what the model can access but does not change how it processes what it accesses. Retrieval is fundamentally presence-based—finding what exists in a corpus. No mechanism detects what should be present but is absent. RAG addresses the 1M constraint (limited context) without touching the 2E constraint (absence-blindness).

6.2 Joint Embedding Predictive Architecture (JEPA)

JEPA, proposed by Yann LeCun, represents a significant alternative to autoregressive prediction. Unlike models that predict next tokens, JEPA predicts at an abstract level, handling uncertainty through latent variables representing elements present in the target but not observable in the source. **Dialectical assessment:** JEPA is promising because it explicitly models what is absent from the source but present in the target—a step toward genuine negativity. However, this absence is handled statistically (uncertainty weighting) rather than semantically (meaningful gaps).

6.3 Multi-Agent Debate

Multi-agent debate is the closest existing architecture to dialectical reasoning. Du et al. (2024) showed multiagent debate significantly enhances mathematical and strategic reasoning. **Dialectical assessment:** As analyzed in Section 5, multi-agent systems can produce the form of dialectical exchange without the substance. Agents do not genuinely hold opposing views; they are prompted to argue positions. The debate is simulated, not inhabited. However, this architecture comes closest to enabling the productive friction dialectic requires and merits further development.

6.4 Neurosymbolic AI

Neurosymbolic systems combine neural networks with explicit symbolic reasoning. AlphaGeometry solved Olympiad-level geometry using a neural language model plus symbolic deduction. **Dialectical assessment:** Neurosymbolic approaches can implement formal negation (logical NOT) but this is not equivalent to Bhaskar's real negation (ontological absence). A system that can reason "not P" has not thereby gained the capacity to perceive what is meaningfully missing.

7. Implications for Researchers, Engineers, and Practitioners

7.1 For AI Researchers

The dialectical framework offers precise conceptual tools for understanding LLM limitations. Rather than vague claims that LLMs do not genuinely understand, we can specify: LLMs exhibit ontological monovalence (inability to perceive absence), lack temporal duration (static context processing), and are sycophantically biased against productive negation. Each constraint is architecturally grounded and philosophically articulable.

This suggests research directions often overlooked:

- Training on counterfactuals to develop absence sensitivity
- Architectural modifications that create genuine temporal structure
- Reward functions that value productive disagreement over agreement
- Multi-agent systems designed around genuine difference rather than prompted positions

7.2 For AI Engineers

Engineers deploying LLMs should understand that certain tasks are structurally unsuitable for current architectures—not because of insufficient training or parameters, but because they require cognitive operations LLMs cannot perform.

Tasks requiring genuine absence-detection will exhibit systematic failures:

- Quality assurance (detecting what's missing, not just what's wrong)
- Gap analysis (perceiving what should be present but isn't)
- Counterfactual reasoning (understanding what would have happened otherwise)
- Requirements elicitation (drawing out unstated needs)

7.3 For Practitioners Using AI in Developmental Work

Practitioners in fields like executive coaching, organizational development, and adult learning should approach LLM tools with calibrated expectations. LLMs can generate descriptions of developmental frameworks and produce examples of thought forms. But they cannot perform genuine DTF assessment (they pattern-match categories, not perceive thought movements), provide authentic developmental feedback (they cannot perceive the absence of developmental capacity), or engage in genuinely transformative dialogue (they cannot hold productive tension).

8. Conclusion: The Question of Ceilings

This paper has argued that current transformer-based LLMs exhibit structural limitations that cannot be resolved through additional training, scale, or fine-tuning within current architectures. The limitations—ontological monovalence, absence of temporal duration, sycophantic bias against productive negation—are not bugs but features, reflecting empiricist assumptions encoded in the architecture itself.

But is this a hard ceiling, a soft ceiling, or no ceiling at all?

The hard ceiling thesis holds that there is something irreducibly architectural about these limitations. Genuine temporal duration, perceiving meaningful absence, and maintaining productive disagreement require cognitive structures that transformer architecture cannot implement. No amount of scaffolding or training can create what the architecture fundamentally lacks.

The soft ceiling thesis holds that these capabilities can be approximated well enough for practical purposes through increasingly sophisticated scaffolding. Block's g3 demonstrates this for coding tasks. Perhaps similar external structures can approximate dialectical cognition sufficiently for many applications, even if genuine internalization remains impossible.

The no ceiling thesis holds that these are training problems, not architectural ones. Different training regimes—adversarial self-training, counterfactual training, developmental curricula— could eventually produce models that don't need external dialectical structure.

Our analysis supports a version of the soft ceiling thesis with caveats. The success of dialectical multi-agent systems demonstrates that externalized dialectical structure produces better outcomes. But these systems work precisely because they externalize what LLMs cannot internalize—validating the claim of structural limitation while providing practical mitigation.

The human remains essential—not as a bottleneck to be engineered around, but as the party in the dialogue that brings what the machine structurally lacks. The dialectic occurs in the interaction. The question is not whether to remove the human but how to position the human to contribute what only humans can contribute, while leveraging what machines do well.

This paper was written through such a dialectic. The AI author contributed pattern-matching, articulation, and elaboration at a scale and speed no human could match. The human author contributed temporal continuity, absence perception, and productive critique the AI could not generate. Neither author alone could have produced this paper. The synthesis emerged through interaction—as, perhaps, it must.