

Adaptive Recursive Convergence and Semantic Turning Points: A Self-Verifying Architecture for Progressive AI Reasoning

Ziping Liu

ziping@gaiaverse.ai

Gaiaverse LTD

Moriba Jah

Gaiaverse LTD

Article

Keywords: Adaptive computation, progressive cognition, dynamic attention, semantic turning point, contraction mappings, token efficiency

Posted Date: July 7th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-6605714/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Adaptive Recursive Convergence and Semantic Turning Points: A Self-Verifying Architecture for Progressive AI Reasoning

ZIPING LIU*, Black Swan Research Group, Gaiaverse Ltd., USA

MORIBA JAH, Black Swan Research Group, Gaiaverse Ltd., USA and Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, USA

Current AI reasoning architectures typically rely on static representational capacities, inherently limiting their ability to adaptively respond when confronted with escalating semantic complexity. Consequently, these models frequently mistake superficial coherence for genuine semantic resolution. To address this fundamental limitation, we introduce the novel paired mechanisms of Adaptive Recursive Convergence (ARC) and Cascading Re-Dimensional Attention (CRA). ARC orchestrates recursive semantic contraction within shared atomic memory, systematically refining representational granularity. In parallel, CRA retroactively evaluates epistemic sufficiency through a rigorously derived confidence score computed via geometric attention measures. When representational saturation occurs—indicating epistemic insufficiency—CRA proactively triggers targeted dimensional escalation, driven entirely by intrinsic semantic detection rather than fixed heuristics. We instantiate and empirically validate ARC/CRA through a publicly available Semantic Turning Point Detector¹, a concrete system capable of segmenting dialogue into epistemically grounded segments by pinpointing structural shifts frequently overlooked by static-dimensional models. This approach demonstrably achieves progressive cognition—dynamically expanding representational scope precisely when necessary, conserving computational resources, and incorporating rigorous self-verification directly into the reasoning process. Ultimately, ARC/CRA reconceptualizes intelligence as a recursive, introspective, and self-adaptive process, continuously capable of recognizing and responding to its own representational limitations, thereby offering a robust blueprint for AI systems to evolve autonomously whenever epistemic certainty breaks down.

CONTENTS

Abstract	1
Contents	1
1 Motivation	2
1.1 Contributions	3
1.2 ARC/CRA as an Operational Occam’s Razor and Epistemic Governor	3
1.3 Related Work	4
2 Methods	5
2.1 Adaptive Recursive Convergence (ARC)	5
2.2 Cascading Re-Dimensional Attention (CRA)	8
3 Results	10
3.1 Semantic Turning-Point Detector	10
3.2 Empirical Validation: Comparative STP Detection Across Diverse Language Models	15
4 Discussion	18
4.1 Beyond Static Dimensionality: The Empirical Case for Adaptive Reasoning Substantiated by ARC/CRA	18

*Corresponding Author.

¹We define a semantic turning point as a location within a dialogue or narrative where the underlying semantic structure undergoes a significant shift, necessitating adaptive representational adjustment. For further information, visit: <https://github.com/gaiaverseltd/semantic-turning-point-detector>.

Authors’ Contact Information: Ziping Liu, ziping@gaiaverse.ai, Black Swan Research Group, Gaiaverse Ltd., Austin, TX, USA; Moriba Jah, moriba@gaiaverse.ai, Black Swan Research Group, Gaiaverse Ltd., Austin, TX, USA and Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX, USA.

4.2	Applications and Forward Trajectories	21
5	Conclusion	22
A	Formalization of the Confidence Score	23
A.1	Properties	23
A.2	Boundary Behavior	24
B	Information-Theoretic Formulation of Complexity Oracle χ	24
B.1	Alternative: Surprisal-Based Formulation	24
B.2	Remarks	24
C	Complexity and Memory Analysis	24
D	Reproducibility Checklist	26
	References	32

1 Motivation

Systems designed for reasoning about open-ended tasks must carefully balance their representational expressivity. Under-parameterization—employing representations too constrained to accurately capture emergent complexities—results in insufficient descriptive power and potentially unreliable inference. Conversely, over-parameterization, characterized by excessively large or unnecessarily complex representational dimensions, introduces prohibitive computational costs and can obscure salient features. Standard machine-learning paradigms, whether fixed-depth transformer architectures [10], static-rank graph embeddings [7], or manually engineered symbolic systems, fundamentally rely upon a *single*, preselected representational dimensionality. This inherent rigidity proves particularly limiting in domains where the intrinsic complexity of a problem progressively unfolds as the reasoning process itself advances. Such domains include multi-step scientific discovery, climate modeling with intricate feedback dynamics, and real-time space-traffic management, all of which necessitate a more fluid and adaptive approach to representation.

To address this critical limitation, we propose a principled two-tier framework designed to **adapt its representational capacity on-demand**, while crucially providing explicit theoretical guarantees of convergence (detailed rigorously in Section 2). This framework is realized through two synergistic core components. The first is **Adaptive Recursive Convergence (ARC)**, which orchestrates multiple concurrent local reasoning processes through a shared *atomic memory*. Leveraging the mathematical foundations of contraction mappings and fixed-point theory [1], ARC formally determines at each recursive iteration whether the current representational dimension possesses adequate complexity to resolve the task at hand. The second component is **Cascading Re-Dimensional Attention (CRA)**. Should ARC detect that the current representation is saturated—meaning its complexity exceeds the capacity of the current dimension—CRA selectively activates a dimension-escalation mechanism. This is not an arbitrary increase in parameters; CRA employs an attention-guided dimensional expansion, analogous to dynamically adding transformer heads, which is explicitly controlled by a discrete, epistemically motivated complexity oracle, χ (formally defined in Section 2.2). Together, ARC and CRA constitute an integrated, closed-loop reasoning architecture: ARC executes *recursive refinement* within a dimension, while CRA facilitates *targeted dimensional augmentation* between dimensions. This synergy yields a powerful form of *progressive cognition*, whereby the system dynamically scales its representational complexity *precisely when warranted* by the epistemic demands of the task, yet remains parsimonious by utilizing simpler abstractions whenever they prove sufficient.

1.1 Contributions

This paper advances the field of adaptive AI reasoning through several principal contributions that collectively offer a novel synthesis of theoretical rigor and practical architectural innovation. A primary contribution is the development of a **contraction-safe recursive framework**. We achieve this by rigorously formalizing Adaptive Recursive Convergence (ARC) as an operator $\Psi : \bar{X} \rightarrow \bar{X}$ on an augmented metric space, and, as established in Theorem 2.4 (Section 2.1), we prove its finite-time convergence to a stable fixed point within an optimal dimension n^* . This provides a robust foundation for computationally efficient and representationally stable adaptive reasoning. Building upon this, our work introduces **an attention-governed dimensional expansion** mechanism via Cascading Re-Dimensional Attention (CRA). This operator utilizes transformer-style QKV soft attention not merely for sequence processing, but for the targeted augmentation of representational dimensionality. Critically, as detailed in Section 2.2, CRA activates this dimensional escalation strictly and only when the complexity classifier χ detects that the current representational capacity is saturated, thereby directly coupling dimensional growth with verified epistemic need.

Furthermore, we establish a **unifying theoretical bridge** that connects symbolic fixed-point recursion with the neural attention mechanisms prevalent in modern transformer architectures [10]. Specifically, we articulate how ARC’s atomic-memory based recursion and CRA’s attention-guided kernel can be interpreted as conceptual duals of self-attention processes, linking foundational symbolic methods rooted in works like Banach’s [1] with contemporary neural network practices, a connection also explored in the broader context of neuro-symbolic AI [3]. Finally, this research proposes a novel **template for uncertainty quantification** inherent within the CRA mechanism. By leveraging the same attention weights that guide dimensional expansion, our framework generates an epistemic *masked-confidence metric*. This metric is designed to quantify the system’s trustworthiness at both intermediate reasoning steps and its final outputs, thereby enabling robust self-verification and creating an introspective reasoning loop vital for the development of reliable AI systems.

1.2 ARC/CRA as an Operational Occam’s Razor and Epistemic Governor

The philosophical principle of Occam’s Razor advocates for selecting explanations that minimize assumptions and complexity while adequately accounting for evidence. The ARC/CRA architecture operationalizes this tenet through a recursive, introspective control loop that actively governs representational complexity. This process begins with **ARC** attempting to solve a given problem within the current representational dimension n , applying contraction mappings over its shared atomic memory to iteratively converge towards a solution. Concurrently, **CRA** performs a retrospective epistemic audit of this reasoning process within dimension n , primarily through its retro-masked confidence signal, C . A critical juncture occurs if the reasoning process becomes saturated, as indicated by the complexity oracle ($\chi(x) = 5$), and if this saturation is coupled with demonstrably low epistemic trust in the generated outputs (e.g., $C < \tau_{\min}$ or a significant decline in confidence $\Delta C < -\tau_{\Delta}$). Under these specific, well-defined conditions, ARC initiates a principled escalation of the reasoning process to dimension $n + 1$.

This dynamic interplay directly implements Occam’s Razor: the system inherently prefers simpler representations (solutions within lower dimensions) and only adds complexity (escalates dimension) when presented with clear, quantifiable evidence of current inadequacy. It starts with minimal assumptions by initializing at a low representational dimension and avoids infinite regress through a finite global depth budget $n_{\max}$ that guarantees eventual termination. More than just a preference for simplicity, ARC/CRA functions as an **Epistemic Razor**. It rigorously demands explicit *evidence of representational inadequacy*, as determined by its internal self-assessment mechanisms, before sanctioning any expansion of its representational capacity. This ensures a systematic and ongoing balance between the expressive power necessary for complex tasks and the maintenance of epistemic soundness and trustworthiness in the reasoning process.

1.3 Related Work

While the aspiration for AI systems with truly adaptive representational complexity is significant, it remains a relatively underexplored domain compared to efforts focused on enhancing fixed-capacity models or varying computational depth within static representations. Nonetheless, ARC/CRA builds upon, and critically distinguishes itself from, several important research trajectories.

1.3.1 *Approaches for Adaptive Computation and Depth: Fixed Dimensionality Constraints.*

Adaptive Computation and Dynamic Architectures. Foundational work on *Adaptive Computation Time (ACT)* by Graves [5] demonstrated that recurrent neural networks can learn to dynamically modulate their number of computational steps per input. This is achieved through a learnable halting mechanism, allowing networks to ‘ponder’ more on complex inputs. Later developments, such as *SkipNet* by Wang et al. [11], introduced dynamic layer skipping in convolutional networks, tailoring computational pathways based on sample difficulty. These approaches, along with others like QADM-Net [8] that adjust depth based on multimodal data quality, achieve notable gains in efficiency. However, as elucidated in their respective works, they primarily operate within a fixed overall representational dimensionality, adjusting the *extent* or *path* of computation rather than the intrinsic *capacity* or *nature* of the representations themselves. ARC/CRA extends this concept by not only allowing for adaptive processing (implicitly, as ARC iterates towards convergence or saturation) but by providing a mechanism through CRA to explicitly manage and expand the dimensionality of the representational space itself when epistemic verification indicates the current scope is insufficient.

Transformers and Depth Adaptivity. The standard Transformer architecture [10], while revolutionary for its parallel processing and attention mechanisms, employs a static number of layers and thus a fixed computational depth. **Universal Transformers (UTs)**, introduced by Dehghani et al. [2], address this by applying a recurrent refinement block, incorporating mechanisms like ACT to dynamically halt processing on a per-position basis, thereby achieving adaptive depth and, under certain conditions, Turing-completeness. Nevertheless, while UTs vary the number of iterative processing steps, they typically do so over representations of a fixed dimensionality within each effective layer. ARC/CRA offers a more fundamental adaptation by decoupling the intra-dimensional recursive refinement managed by ARC from the principled, discrete inter-dimensional expansion orchestrated by CRA. This allows ARC/CRA to modify not just the processing depth but the core expressive power of its representations, guided by explicit epistemic trust signals and a formal complexity oracle, a distinction from the primarily depth-adaptive nature of UTs.

1.3.2 *Reasoning Frameworks, Symbolic-Neural Hybrids, and the Motivation for Epistemic Dimensional Adaptivity.*

Multi-Step Reasoning and Agentic AI Frameworks. Modern agentic frameworks, including those employing ReAct as proposed by Yao et al. [13] or Chain-of-Thought (CoT) prompting as explored by Wei et al. [12], have demonstrated significant capabilities in decomposing tasks and orchestrating multi-step reasoning, often by leveraging the generative and inferential power of large language models. These systems excel at interleaving reasoning traces with actions or generating intermediate thoughts to guide problem-solving. However, their mechanisms for escalating reasoning complexity or verifying the fidelity of intermediate conclusions frequently rely on the LLM’s implicit, often opaque, internal confidence or on heuristic measures. They generally lack formal guarantees of convergence within a reasoning stage or transparent, mathematically grounded criteria for deciding when and how to deepen or expand the representational basis of their reasoning. ARC/CRA addresses this gap by introducing an explicit, contraction-based mechanism for intra-dimensional convergence (ARC) coupled with a CRA-driven confidence assessment and expansion policy. This provides a more formal and verifiable theoretical foundation for self-regulated escalation and epistemic validation within complex reasoning sequences, potentially complementing the procedural strengths of agentic frameworks.

Neuro-Symbolic and Probabilistic Methods. The endeavor to integrate the structured rigor of symbolic reasoning, such as fixed-point theory, with the adaptive learning capabilities of neural architectures is a central theme in neuro-symbolic AI [3]. Similarly, probabilistic programming languages like Church, as described by Goodman et al. [4], allow for the description of stochastic generative processes and support inference over complex models. While these fields strive to unify interpretability, formal semantics, and learning, they have seldom incorporated mechanisms for the adaptive scaling of representational dimensionality that is directly governed by runtime epistemic signals. ARC/CRA bridges this by grounding its recursive refinement processes (ARC) in the formal guarantees of contraction mappings, akin to fixed-point iterations, while using attention-derived confidence signals (CRA) to drive flexible, yet principled, dimensional adaptivity. This offers a novel path towards integrating symbolic stability with neural adaptability in a dynamically scalable manner.

Long-Context Fragility of Large Models. Recent evaluations, exemplified by the NoLiMa benchmark [6], underscore a critical challenge for many LLMs: a significant degradation in performance with increasing context lengths. This ‘context-length paradox’ highlights the limitations of simply scaling context windows via brute-force. ARC/CRA offers a potential alternative by enabling intelligent, selective processing. Instead of uniformly processing vast contexts, ARC/CRA’s mechanisms for identifying semantic turning points and varying representational depth and dimensionality allow it to focus computational resources and enhance representational richness precisely where semantic complexity demands it, offering a more sustainable approach to handling long-form inputs.

In essence, ARC/CRA distinguishes itself by operationalizing a self-verifying, dimensionally adaptive reasoning process. By systematically integrating formal convergence guarantees derived from fixed-point theory with dynamic epistemic verification through novel attention-based mechanisms, our framework advances adaptive AI reasoning beyond existing approaches. It enables a form of progressive cognition that is precisely tailored to evolving complexity demands, ensuring both robustness and resource efficiency in navigating the intricate landscape of open-ended problems.

2 Methods

Section 2.1 develops the ARC formalism, while Section 2.2 details CRA and shows how the two mechanisms interlock to keep resource growth bounded.

Following this section is *Results* (§3), which provides empirical validation and an explanation of one possible implementation of Section 2. This implementation, known as the Semantic Turning Point defined fully in Section 3.1, is already referenced with its code repository at <https://github.com/gaiaverseltd/semantic-turning-point-detector>. The repository offers a valuable tool that efficiently and rapidly addresses a qualitative and linguistic problem, regardless of the LLM model used.

2.1 Adaptive Recursive Convergence (ARC)

Take-away. ARC is the *executive cortex* of our architecture. It (i) orchestrates many reasoning threads through a **shared atomic memory**, (ii) guarantees that each local sweep is a *true contraction*, (iii) decides—via a lightweight complexity oracle χ —whether the current abstraction is *still adequate*, and (iv) enforces a *depth-limited* escalation rule that makes divergence *provably impossible*. CRA (§2.2) only runs when ARC certifies that another semantic dimension is strictly necessary.

2.1.1 Formal Setting. Let (X, ρ) be a complete metric space encoding the *global configuration*. Suppose $m \in \mathbb{N}$ concurrent agents whose private state factorizes as

$$x = (x_1, \dots, x_m) \in S_1 \times \dots \times S_m \subseteq X.$$

Definition 2.1 (Augmented Configuration). The working space of ARC is $\bar{X} = X \times \mathbb{N}$, where (x, n) records the configuration x and the *reasoning dimension* $n \geq 0$ —the number of representational expansions performed so far.

2.1.2 *Atomic Memory and Local Sweeps.* At iteration k each agent updates

$$x_i^{(k+1)} = f_i(x_i^{(k)}, A^{(k)}), \quad (\text{SWEEP})$$

where the **atomic memory tensor**

$$A^{(k)} = \Phi(x_1^{(k)}, \dots, x_m^{(k)}) \in \mathbb{A}$$

is produced by an aggregator

$$\Phi : X^m \rightarrow \mathbb{A}.$$

A complete sweep $F(x) = (f_1(x_1, A), \dots, f_m(x_m, A))$ is applied synchronously.

Design principle. Every $f_i(\cdot, A)$ is either a strict ρ contraction or decreases a Lyapunov-like error functional $\mathcal{E}_i$. Hence a dimension that is *expressive enough* must converge by Banach’s fixed-point theorem.

2.1.3 *Self-Verification, Cycle Control, and Memory Re-Use.* ARC inserts two micro-kernels *between* sweeps:

- i **Cycle sentinel.** If $(x_i^{(k)}, A^{(k)})$ has appeared before, skip the update—guarantees acyclic progress.
- ii **Frozen-error check.** If $\mathcal{E}_i(x_i^{(k)}) < \varepsilon_i$, hold x_i fixed; other agents may still evolve.

All intermediate artifacts are cached in atomic memory, so any later sub-query that rediscovers an already solved substructure is served *from cache at zero cost*. Together, these rules yield *monotone descent of the global residual* without redundant computation.

2.1.4 *Escalation Trigger: The Complexity Oracle χ .* ARC uses a discrete classifier, the **complexity oracle** $\chi : X \rightarrow \{1, 2, 3, 4, 5\}$, to assess the difficulty of reaching convergence *within the current dimension* n :

$\chi(x)$	Label	Operational Meaning
1	Trivial	Convergence likely in one sweep.
2–4	Refinable	Convergence expected within a few sweeps.
5	Saturated	The expressive capacity of dimension n is insufficient; the contraction gap cannot close.

Implementations can realize χ using various metrics, such as the variance of CRA’s masked-confidence scores (§2.2), residual entropy, or the spectral radius of the Jacobian DF.

Remark 2.2. *The classifier χ can be a lightweight rule-based functional or a trainable module. Future work might involve refining χ dynamically using reinforcement learning feedback, allowing the system to learn optimal predictors of contraction failure or epistemic drift based on runtime experience.*

2.1.5 *Depth-Limited Escalation Policy.* Let $n_{\max} \in \mathbb{N}$ be a *global depth budget*. Denote by Conf_{out} the retro-masked confidence produced by CRA (see §2.2). ARC decrees

$$n_{\text{next}} = \begin{cases} n, & \chi(x) < 5 \text{ and } \text{Conf}_{\text{out}} \geq \tau_{\text{good}}, \\ n + 1, & \chi(x) = 5 \text{ and } n < n_{\max}, \\ \text{halt}, & \text{otherwise.} \end{cases} \quad (\text{ESCALATE})$$

Thus every jump *consumes* one unit of the finite budget $n_{\max}$ and is triggered *only if* the confidence signal certifies that dimension n is inadequate. The stack therefore converges in at most $n_{\max}$ hops.

Remark 2.3. *Dimensional escalation is tightly controlled by a global depth budget $n_{\max} \in \mathbb{N}$ and the confidence score C_{out} produced by CRA (§2.2).*

2.1.6 Global Dynamics, Convergence, and System Role. The local sweeps and escalation policy combine to define the system’s overall behavior and theoretical guarantees.

The Global Transition Operator Ψ . Both intra-dimensional sweeps and potential inter-dimensional escalations, is captured by the global operator:

$$\Psi(x, n) = \begin{cases} (F(x), n) & \text{if ESCALATE dictates staying at dimension } n \\ (F(x), n + 1) & \text{if ESCALATE dictates jumping to } n + 1 \\ (x, n) & \text{if ESCALATE dictates halting} \end{cases}$$

Iterating Ψ drives the system through its adaptive reasoning process.

Theorem 2.4 (Finite-Time Convergence with Depth Budget). *Let each local update $f_i(\cdot, A)$ be a ρ -contraction for any fixed memory A , let the complexity oracle χ be non-increasing under F , and let the maximum depth $n_{\max}$ be finite. Then starting from any $(x^{(0)}, 0)$, repeated application of Ψ terminates in a finite number of steps at some (x^*, n^*) satisfying the ESCALATE halting conditions: either*

$$\chi(x^*) < 5 \quad \text{and} \quad C_{\text{out}}(x^*) \geq \tau_{\text{good}},$$

or

$$n^* = n_{\max}.$$

PROOF SKETCH. Within a fixed dimension n , if $\chi(x) < 5$, then F is contractive and converges to a unique fixed point by Banach’s theorem [1]. If instead $\chi(x) = 5$, the ESCALATE rule bumps $n \rightarrow n + 1$ (while $n < n_{\max}$). Since n can increase at most $n_{\max}$ times and each dimension requires only finitely many sweeps to either converge or trigger escalation, the process must halt after a finite sequence of sweeps and escalations at one of the stated stopping criteria $\square$

2.1.7 Interpretive Lens: ARC’s Role and Hand-Off to CRA. From an engineering perspective, ARC provides robust solutions to several critical challenges in complex reasoning systems:

1. **Concurrency Control:** Synchronized sweeps using atomic memory prevent race conditions and ensure consistent state updates across agents.
2. **Computational Efficiency:** Atomic memory caching eliminates redundant computations by reusing previously computed results for recurring subproblems.
3. **Algorithmic Soundness:** The combination of local contractions and cycle sentinels guarantees termination within a dimension and prevents runaway loops.
4. **Adaptive Adequacy:** The complexity oracle χ and the CRA confidence score C_{out} form a *self-awareness loop*, enabling the system to recognize when its current representational capacity is insufficient.

When χ indicates saturation but $n < n_{\max}$, ARC halts further contraction—since stability is guaranteed by Banach’s theorem—and delegates to CRA to instantiate the next-dimensional embedding [1]. Ergo, ARC manages convergence and resource allocation, while CRA enables greater representational capacity to encapsulate the context or environment of the task or problem. Concordantly, this process is illustrated in later sections, for example, within the context of a conversational dialogue [9] through the mechanism of the Semantic Turning Point (see Section 3.1).

2.2 Cascading Re-Dimensional Attention (CRA)

Essence. CRA is the *meta-cognitive complement* to ARC. ARC orchestrates *how long* the system should iterate inside a semantic layer; CRA decides *whether that layer can still be trusted*. It does so by *reversing* the role of attention masks: instead of hiding context *before* inference (as in classical QKV), CRA uncovers a mask *after* inference that quantifies the epistemic soundness of every reasoning fragment. That retro-mask collapses to a single scalar $C \in [0, 1]$ that functions simultaneously as

- a **certainty oracle** — high when reasoning is coherent and schema-grounded, low when it drifts;
- an **escalation trigger** — fed back into ARC to decide if we must jump from dimension n to $n+1$.

2.2.1 The Retro-Mask: An Epistemic Mirror. CRA’s core innovation is the inversion of traditional attention masking. Instead of filtering out context, it post-evaluates reasoning fragments, creating a diagnostic mirror reflecting epistemic quality.

Contrast with Forward Masking. In standard Transformers, given query q_i , keys k_j , and values v_j , the forward mask $M_{ij} \in \{0, -\infty\}$ *preemptively* restricts information flow before the softmax operation:

$$\alpha_{ij} = \frac{\exp(q_i \cdot k_j + M_{ij})}{\sum_{j'} \exp(q_i \cdot k_{j'} + M_{ij'})}, \quad \sum_j \alpha_{ij} v_j.$$

This mask enforces constraints like causality or padding, shaping *what information influences* the output.

The CRA Retro-Masking Process. CRA operates differently:

- Allow Reasoning:** First, all sub-analyses A_i (e.g., from different attention heads or recursive steps) are allowed to complete without prior masking.
- Post-Hoc Evaluation:** Only then does CRA compute metrics for each fragment A_i :

$$C_i = \cos(A_i, A_{\text{agg}}), \quad G_i = \cos(A_i, E_d),$$

$$M_i = \left(\frac{C_i + 1}{2} \right) \left(\frac{G_i + 1}{2} \right) \in [0, 1].$$

Here, $A_{\text{agg}} = \sum_j w_j A_j$ represents an attention-weighted centroid (overall consensus) of fragments, and E_d is the domain-schema embedding (ground truth/context).

- Form the Retro-Mask:** The resulting vector $\mathbf{M} = (M_1, \dots, M_N)$ acts as a *revealed* or *diagnostic* mask. It rates, **after-the-fact**, the internal coherence (C_i) and external grounding (G_i) of each reasoning step.

	Forward Mask (Transformers)	Retro Mask (CRA)
Applied when	<i>Before</i> attention softmax	<i>After</i> reasoning completes
Purpose	Restrict visible context	Rate epistemic fidelity
Effect on A_f	Causal/padding compliance	<i>None directly</i> (guides control)
Information	Syntactic/Structural Validity	Semantic Certainty/Coherence

This fundamental inversion turns the mask’s role from a *context filter* into an *epistemic mirror*. Because this mirror arises from the model’s own internal geometry (cosine similarities), it yields a *semantic* estimate of certainty, distinct from purely statistical measures (like logit entropy).

2.2.2 Confidence Quantification and Escalation Control. The retro-mask $\mathbf{M}$ provides the raw data, which is then distilled into a scalar confidence score C that drives ARC’s dimensional escalation decisions.

Computing the Confidence Scalar C . The confidence score integrates internal process harmony (C_{in}) and final output alignment (C_{out}):

$$C_{\text{in}} = \frac{1}{N} \sum_i M_i, \quad C_{\text{out}} = \left(\frac{\cos(A_f, A_{\text{agg}}) + 1}{2} \right) \left(\frac{\cos(A_f, E_d) + 1}{2} \right),$$

where A_f is the final synthesized output vector. These components are blended, typically using the golden ratio $\lambda \approx 0.618$ to slightly favor the output quality:

$$C = \lambda C_{\text{out}} + (1 - \lambda) C_{\text{in}}.$$

Interpretation:

- C_{in} : Reflects the consistency and grounding of the reasoning steps taken (*journey quality*).
- C_{out} : Reflects the coherence and grounding of the final result (*destination quality*).
- C : Provides a holistic confidence measure.

Robustness: If only one fragment remains ($N=1$), C_{in} simplifies to a cosine similarity term, preventing failure and still providing a lower-bound signal.

The Saturation Test: Triggering Dimensional Escalation. ARC uses C to determine if the current reasoning dimension n is **saturated**. Saturation occurs if internal coherence is too low or if the synthesis process degrades trust:

$$\text{Layer } n \text{ saturated} \iff C_{\text{in}} < \tau_{\min} \vee (C_{\text{out}} - C_{\text{in}}) < -\tau_{\Delta}, \quad (\text{SAT})$$

where $0 < \tau_{\min}, \tau_{\Delta} < 1$ are sensitivity thresholds.

- Condition 1 ($C_{\text{in}} < \tau_{\min}$): Reasoning fragments lacked internal agreement or drifted from the schema.
- Condition 2 ($C_{\text{out}} - C_{\text{in}} < -\tau_{\Delta}$): The final output is significantly less trustworthy than the intermediate steps.

If (SAT) is true, ARC initiates dimensional escalation $(x, n) \mapsto (x, n+1)$. CRA then generates a new, potentially higher-capacity, reasoning context (e.g., a new QKV stack) for dimension $n + 1$.

2.2.3 Operational Dynamics and System Impact. CRA functions as the sensing mechanism within ARC's adaptive control loop, enabling the system to intelligently manage its representational capacity based on measured confidence.

Algorithmic Flow within a Dimension. The interplay within a single dimension n follows these steps:

- (1) **Reasoning Sweep:** Execute the core reasoning process (e.g., multi-head attention) to produce intermediate fragments $A_{1:N}$ and a final output A_f .
- (2) **Confidence Assessment:** Compute the retro-mask $\mathbf{M}$ and derive the confidence metrics ($C_{\text{in}}, C_{\text{out}}, \Delta C$).
- (3) **Saturation Check:** Evaluate the (SAT) condition using the confidence scores.
- (4) **Control Decision:**
 - If (SAT) is true and $n < n_{\max}$: Signal ARC to escalate to dimension $n + 1$.
 - If (SAT) is false and $C_{\text{out}} < 1 - \varepsilon$: Repeat the reasoning sweep (refine within dimension n).
 - If (SAT) is false and $C_{\text{out}} \geq 1 - \varepsilon$: Terminate successfully within dimension n .
 - If (SAT) is true and $n = n_{\max}$: Terminate due to reaching depth budget.
 - Also terminate if resource limits (e.g., iterations per layer) are exceeded.

Guarantees and Budget-Aware Termination. The global depth budget $n_{\max}$ enforced by ARC (§2.1.5) ensures CRA cannot trigger infinite dimensional escalations. Each jump consumes one unit of this budget. Furthermore, mechanisms like annealing the thresholds ($\tau_{\min}, \tau_{\Delta}$) or setting iteration limits within each layer guarantee that the process terminates in finite time, either by achieving sufficient confidence or by exhausting the allocated resources ($n_{\max}$ or iteration limits).

Closing the Attention-Confidence Loop: Overall Impact. CRA fundamentally changes the role of attention-related metrics by feeding them back into the system’s control logic. The confidence score C dictates the system’s trajectory:

$$C \begin{cases} \text{High} & \Rightarrow \text{ARC continues refinement in dimension } n. \\ \text{Low / Degraded} & \Rightarrow \text{ARC triggers escalation to dimension } n+1. \end{cases}$$

Summarizing the paradigm shift:

- **Forward Masks:** Control *what input is processed*.
- **Retro Masks (CRA):** Assess *how trustworthy the output was*.

By linking the retro-mask’s assessment back to ARC’s dimensional control, CRA creates a closed loop that dynamically balances *reasoning depth* with *epistemic certainty*. This enables a system that adapts its representational capacity in response to the perceived quality of its own reasoning:

$$\text{Reason} \xrightarrow{\text{CRA}} \text{Measure Trust} \xrightarrow{\text{ARC}} \text{Adapt Capacity}.$$

3 Results

3.1 Semantic Turning-Point Detector

Having delineated the theoretical architecture of Adaptive Recursive Convergence (ARC) for principled intra-dimensional reasoning (Section 2.1) and Cascading Re-Dimensional Attention (CRA) for epistemically-grounded dimensional adaptation (Section 2.2), this section presents their Empirical Realization and validation via an implementation as already mentioned by name, the **Semantic Turning-Point (STP) Detector**. The STP Detector is not *a mere downstream application* of ARC/CRA principles; rather, it serves as a live, unsupervised adaptive system that fully instantiates the ARC/CRA operational loop. Its design and observed performance offer concrete, verifiable demonstrations of ARC’s core *mechanisms*—*contraction-driven* refinement, CRA’s *confidence-calibrated dimensional escalation*, and the system’s overall dynamic responsiveness to evolving semantic complexity. The empirical findings obtained from applying the STP Detector provide direct validation for the core claims and operational dynamics of the ARC/CRA framework, which are discussed in greater detail in Section 2, Subsection 3.2.

To ensure complete transparency, facilitate rigorous reproducibility, and encourage broader academic and developmental engagement, the entire STP Detector implementation is open-sourced. The complete TypeScript source code, which embodies the ARC/CRA logic, is available in our GitHub repository². This repository also includes thorough documentation, illustrative usage examples, and logged outputs from benchmark runs. Furthermore, a readily integrable *NPM* package³ is provided, enabling researchers and developers to apply and experiment with the ARC/CRA-based STP Detector in diverse contexts with minimal setup. These publicly accessible resources constitute direct and verifiable evidence of ARC/CRA’s practical capabilities and the theoretical claims presented throughout this paper, addressing the need for tangible substantiation of our framework.

²See the GitHub repository at <https://github.com/gaiaverseltd/semantic-turning-point-detector>.

³Access the NPM package at <https://www.npmjs.com/package/@gaiaverse/semantic-turning-point-detector>.

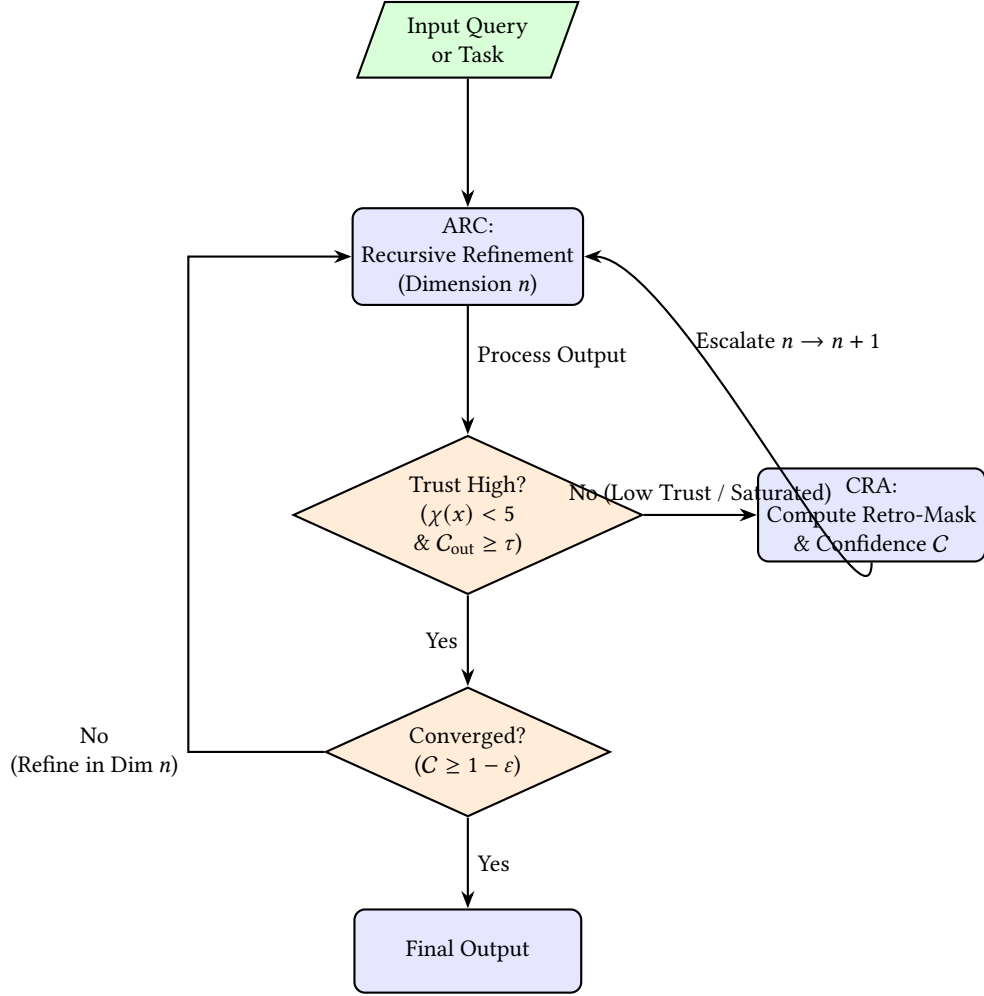


Fig. 1. Illustrates the integrated ARC/CRA flow

To emphasize the practical utility of the code—which is designed to function as a standalone package rather than merely fulfilling a requirement for this paper—the STP detector is compatible with any LLM API endpoint that adheres to the OpenAI format. This includes, but is not limited to, endpoints provided by Ollama⁴, LMStudio⁵, vLLM⁶, and Text Generation WebUI⁷.

Definition 3.1 (Semantic Turning Point (STP)). Within a sequence of message embeddings $\{M_t\}_{t=1}^T$ in a given n -dimensional space $\mathbb{R}^n$, representing a dialogue or narrative, a specific time index $t^* \in \{2, \dots, T-1\}$ is identified

⁴<https://github.com/ollama/ollama>

⁵<https://lmstudio.ai/>; primarily aimed at running LLMs on Apple Silicon Macs and MacBooks

⁶<https://github.com/vllm-project/vllm>; primarily designed for running LLMs via PyTorch as efficiently as possible on modern GPUs

⁷<https://github.com/oobabooga/text-generation-webui>; a robust alternative to Ollama that supports loading LLMs through various methods and provides tools for fine-tuning models loaded via <https://huggingface.co/docs/transformers>

as a **Semantic Turning Point (STP)**. The primary condition for an STP is the presence of **High Local Semantic Curvature**, which signifies a substantial second-order directional change in the embedding space. This is quantified as $\kappa(t^*) = \|M_{t^*-1} - 2M_{t^*} + M_{t^*+1}\| > \delta_\kappa$, for a given sensitivity threshold $\delta_\kappa > 0$. For t^* to be confirmed as a valid STP, this primary curvature must be accompanied by at least one of two secondary epistemic indicators derived from the CRA mechanism:

- (b) A **Confidence Gap Jump**, where the CRA retro-masked confidence score $C(t)$ (as defined in Section 2.2.2) demonstrates a significant decrease in trust between the analytical process and its synthesized output, formally $\Delta C(t^*) = C_{\text{out}}(t^*) - C_{\text{in}}(t^*) < -\tau_\Delta$.
- (c) A **Contextual Shift Divergence**, where the cosine distance between aggregated embeddings of local segments preceding and succeeding t^* changes sharply, i.e., $D_{\cos}(t^*) = 1 - \cos(\bar{M}_{[t^*-k:t^*-1]}, \bar{M}_{[t^*+1:t^*+k]}) > \theta_{\text{div}}$, for a specified window size k and divergence threshold θ_{div} .

A point t^* is thus considered a valid STP if it satisfies condition (a) and at least one of conditions (b) or (c), indicating a structurally significant and epistemically notable shift in the semantic landscape.

3.1.1 Epistemic Realization: The ARC/CRA Operational Cycle in STP Detection. The Semantic Turning-Point (STP) Detector serves as an empirical instantiation that fully operationalizes the Adaptive Recursive Convergence (ARC) and Cascading Re-Dimensional Attention (CRA) frameworks. In practice, the detection cycle unfolds through a structured yet flexible series of epistemically-grounded steps, systematically leveraging the complementary strengths of ARC and CRA to identify semantic shifts within dialogues.

The detection process initiates at a foundational representational dimension ($n = 0$) through a procedure termed **Semantic Decomposition**. Rather than imposing external heuristics, ARC dynamically applies contraction mapping principles to each message embedding within the input dialogue. Local peaks in the cosine similarity shifts between consecutive message embeddings naturally emerge as candidate turning points. This approach fundamentally treats semantic segmentation as an organic discovery process, capturing intrinsic dialogue structure rather than forcing externally defined segment boundaries.

Once candidate segments are identified at this initial dimension, each segment undergoes a rigorous **Expressivity Saturation** assessment via CRA. Here, CRA serves as an intrinsic auditor, evaluating whether the current dimensional representation is sufficient to capture the complexity contained within a given segment. This evaluation is quantified by a complexity oracle χ , calculated through the product of semantic shift magnitude and context entropy. A threshold $\chi = 5$ explicitly signals representational saturation, clearly indicating that the current dimension can no longer adequately represent the underlying semantic complexity.

Upon detection of representational insufficiency, ARC responds with a strategically precise action of **Targeted Dimensional Escalation**. This escalation is not indiscriminate or heuristic-driven; rather, it selectively elevates the representational dimensionality strictly for those segments identified by CRA as saturated. Thus, the representational expansion from dimension n to $n + 1$ remains economical and contextually justified, applied solely in situations where deeper semantic complexity demonstrably demands higher expressivity. This targeted approach ensures computational efficiency, avoiding unnecessary resource consumption.

The cycle concludes by returning to CRA for an explicit **Confidence and Convergence** verification step. CRA outputs two critical epistemic measures: a retro-mask $\mathbf{M}$ —marking segments that required dimensional escalation—and a scalar confidence score $C \in [0, 1]$. This confidence score quantifies the epistemic reliability and internal coherence of the identified semantic boundaries. ARC utilizes these metrics to make informed decisions to either finalize the segmentation or refine it further, contingent on the achieved confidence levels. Only when all candidate boundaries fall within epistemically justified thresholds—ensuring each segmentation decision is robustly validated by the system’s internal standards—does the adaptive reasoning loop terminate. This comprehensive operational cycle exemplifies a rigorous, self-verifying procedure that actively ensures representational adequacy through epistemic reflection and controlled escalation.

3.1.2 Prioritizing Expressivity and Intrinsic Auditing over External Correctness. Central to the ARC/CRA architecture, as empirically demonstrated through the STP Detector, is an explicit prioritization of representational expressivity and semantic coherence over alignment with any externally imposed definitions of correctness. Rather than attempting to match predefined or gold-standard boundaries, the STP Detector partitions dialogues primarily on the grounds of internal semantic consistency. This foundational design principle ensures segmentation emerges from the intrinsic structural and thematic shifts inherent to the data rather than from externally defined notions.

Further, the confidence score C generated by CRA operates fundamentally as an intrinsic audit rather than an external validation measure. Derived from the internal geometric consistency of embeddings, this confidence score offers the system an epistemically grounded, self-referential assessment of whether the current representation adequately captures observed semantic transitions. Thus, CRA provides an epistemic mirror that objectively reflects the internal coherence and dimensional sufficiency of the segmentation, independent of external criteria or annotations.

Each dimensional escalation triggered by ARC in response to CRA’s saturation signals embodies an act of epistemic self-reflection. The system effectively ‘learns’ its representational limitations and incrementally extends its expressive capacity only where demonstrably necessary—thereby embodying adaptive introspection within the reasoning process. This mechanism ensures that expansions in representational complexity are not arbitrary but epistemically substantiated.

As a direct consequence of these principles, the STP Detector’s termination criteria are driven purely by semantic coherence rather than artificial token limits or iteration counts. The adaptive segmentation process halts precisely when no unresolved complexities or semantic tensions exceed the dynamically adjusted expressivity threshold. This ensures the dialogue’s narrative structure is thoroughly and epistemically resolved by the representation employed. This operational philosophy, rigorously implemented and validated through the publicly available STP Detector codebase hosted on GitHub and distributed via NPM, underscores the ARC/CRA framework’s distinctive approach: semantic segmentation as a fundamentally epistemic, internally justified process rather than externally benchmarked alignment.

3.1.3 The First Complete ARC/CRA Loop in Action: Demonstrating Operational Principles. The Semantic Turning-Point (STP) Detector, openly available through our GitHub repository⁸ and the corresponding NPM package⁹, constitutes a significant advancement beyond mere theoretical conceptualization. It represents the *first fully realized empirical demonstration* of the combined principles of Adaptive Recursive Convergence (ARC) and Cascading Re-dimensional Attention (CRA) operating seamlessly in a closed-loop manner. By applying these principles to complex, real-world dialogue data, the STP Detector serves not only as a proof-of-concept for ARC/CRA but concretely demonstrates several critical operational principles inherent to the architecture.

A core feature vividly demonstrated by this operational instantiation is its robust form of **closed-loop reasoning**. Within each iterative step, ARC systematically employs its contraction mapping principles to assess semantic coherence across consecutive messages, identifying semantically stable segments. Concurrently, the CRA component rigorously audits and verifies representational sufficiency, producing an intrinsic confidence assessment at each segment boundary. This confidence score directly informs ARC’s strategic decisions: either maintaining current representational dimensionality for refinement or selectively escalating the dimension only when representational inadequacy is epistemically substantiated. Thus, the Detector embodies a dynamic feedback mechanism, continuously adapting the analytical granularity in response to the semantic complexity presented by the dialogue.

⁸See the GitHub repository at <https://github.com/gaiaverseltd/semantic-turning-point-detector>.

⁹Access the NPM package at <https://www.npmjs.com/package/@gaiaverse/semantic-turning-point-detector>.

A second, equally pivotal operational characteristic illuminated by this implementation is the genuinely **self-organizing structural adaptation**. Unlike conventional semantic segmentation models, which typically depend on externally imposed rules or annotated labels, the STP Detector autonomously adapts its internal representational framework strictly according to the intrinsic semantic patterns that organically emerge from the input data. Rather than imposing arbitrary global dimensional increases, the ARC-triggered representational expansions occur selectively, guided explicitly by epistemic evaluation from CRA. This ensures that representational resources are strategically and contextually allocated precisely where semantic complexities demand richer expressivity, enabling a parsimonious yet highly adaptive analytic process.

Furthermore, the operational performance of ARC/CRA as instantiated through the STP Detector reveals a pronounced **model-agnostic robustness**. Extensive comparative empirical results presented in Section 3.2 and summarized in Table 1 validate this characteristic clearly. Remarkably, even relatively modestly sized language models such as GPT-4.1-Nano and Qwen3-0.6B achieve coherent, high-confidence semantic segmentation when integrated within the ARC/CRA framework. Conversely, applying the same rigorous epistemic auditing procedures to larger models like GPT-4.1 and Gemini Flash 2.5 elucidates subtle interpretative biases and distinctive structural segmentation patterns inherent in their design. Hence, ARC/CRA acts not merely as an analytic method but also as an invaluable epistemic lens through which model-specific interpretive behaviors become systematically visible, verifiable, and quantifiable.

In this context, the publicly available implementation of the STP Detector signifies a critical **conceptual milestone** within the field of adaptive and epistemically grounded reasoning architectures. By concretely embodying the interplay of adaptive recursion and intrinsic confidence evaluation, the implementation empirically demonstrates the capability of ARC/CRA to uncover nuanced narrative structures reliably, outperforming traditional single-dimensional or fixed-capacity models. This advancement highlights the ARC/CRA architecture’s broader potential as an epistemically self-aware, introspective, and dynamically scalable reasoning framework.

Table 1. Comparative Metrics for Semantic Turning-Point Segmentation on a 6,126-token Multi-Turn Dialogue. Metrics include Number of Turning Points (#TPs), Mean, Median, Min/Max Segment Length in turns (ℓ), Average Confidence ($\bar{c}$), and Processing Time. *Local hardware performance shown for Qwen models; Gemini Flash 2.5 performance via OpenRouter API.

Model	#TPs	Mean ℓ	Median ℓ	Min ℓ	Max ℓ	$\bar{c}$	Time (MM:SS)
GPT-4.1	16	6.0	3	2	33	0.51	06:06
GPT-4.1-Nano	13	12.1	6	2	72	0.43	01:15 *
Qwen3-1.7B	16	10.0	4.5	2	72	0.49	51:26*
Qwen3-0.6B	15	10.5	4	2	62	0.45	04:28*
Gemini Flash 2.5 (Thinking)	16	3.6	2	2	15	0.60	30:00*

Timings for **Qwen3-1.7B** (see the Qwen3-1.7B library at <https://ollama.com/library/qwen3:1.7b>) and **Qwen3-0.6B** (see the Qwen3-0.6B library at <https://ollama.com/library/qwen3:0.6b>) reflect local hardware performance: **Qwen3-1.7B** was executed on a CPU-only server (*Intel i9-12900H*) utilizing *Ollama’s SYCL backend* (see the *SYCL backend documentation at <https://github.com/ggml-org/llama.cpp/blob/master/docs/backend/SYCL.md>*) for heterogeneous computation, whereas **Qwen3-0.6B** ran on a legacy NVIDIA Turing-architecture GPU (*Titan RTX, 24GB VRAM*) via *Ollama’s CUDA backend* (see the *CUDA backend documentation at <https://github.com/ggml-org/llama.cpp/blob/master/docs/build.md#cuda>*). These results critically demonstrate that **locally deployed** models, when orchestrated by ARC/CRA, can deliver competitive, near-real-time semantic segmentation. This provides a practical, low-latency, and potentially private alternative to relying on **commercial** LLM APIs. The **Gemini Flash 2.5** variant utilized is one that incorporates an explicit *thinking* step, accessed via the **OpenRouter API** (see the OpenRouter API documentation at <https://openrouter.ai/google/gemini-2.5-flash-preview:thinking/api>).

3.2 Empirical Validation: Comparative STP Detection Across Diverse Language Models

To rigorously validate the performance and capabilities of the Semantic Turning-Point (STP) Detector within the ARC/CRA framework, we conducted an extensive comparative empirical analysis of semantic segmentation across diverse, state-of-the-art language models. Our analysis specifically involved five distinct model architectures: **GPT-4.1**, its smaller and computationally efficient variant **GPT-4.1-Nano**, open-source models **Qwen3-1.7B** and **Qwen3-0.6B**, and the API-based **Gemini 2.5 Flash (Thinking variant)**. The evaluation dataset consisted of a complex, multi-turn dialogue comprising 154 conversational turns, totaling 6,126 tokens. This dataset, selected for its intricate narrative structure and subtle semantic turning points, provided an ideal environment to systematically examine how each language model, mediated through the adaptive ARC/CRA loop, structurally segments dialogue and identifies semantic turning points. This multi-turn dialogue is drawn from August Strindberg’s one-act masterpiece, *Pariah or Outcast* [9].

Pariah or Outcast is an exceptional testbed for this analysis due to its remarkable psychological depth, tightly woven structure, and the density of its semantic inflection points. The play unfolds as a single, uninterrupted conversation between two middle-aged men—Mr. X, an archeologist, and Mr. Y, a traveler from America—set in the confined space of a country house. This minimal cast and setting strip away narrative distractions, focusing the reader on the evolving psychological and ethical landscape. The dialogue is a masterclass in subtext, ambiguity, and shifting power dynamics: what begins as small talk about the weather and daily affairs quickly escalates into confessions of crime, existential self-doubt, and mutual moral interrogation.

Crucially, the play is punctuated by sharp, high-stakes semantic turning points—moments where the conversation pivots from surface-level exchanges to profound revelations about guilt, justice, and the nature of virtue. For example, Mr. X’s internal struggle with the temptation to steal gold, juxtaposed with Mr. Y’s reflections on necessity and criminality, creates repeated inflection points that challenge both characters’ self-understanding and the audience’s moral intuitions. These are not merely plot twists, but deep epistemic and semantic shifts—precisely the kind of transitions that ARC/CRA is designed to detect and reason about.

The dialogue’s structure is further enriched by Strindberg’s use of mirroring and reversal. As the play progresses, the two men’s roles and moral positions blur and invert: the accuser becomes the accused, the confessor becomes the judge, and the boundaries between guilt and innocence, strength and weakness, are repeatedly renegotiated. Consider, for instance, the moment when Mr. Y, himself a forger, probes Mr. X’s conscience: “*Vanity, ambition, could gain pardon where need could not?*” This line encapsulates the play’s central moral ambiguity and serves as a quintessential semantic turning point, where the conversation shifts from abstract theorizing to personal reckoning.

By leveraging this play, we demonstrate how the ARC/CRA framework can uncover nuanced narrative structures that static models often miss. The STP Detector successfully identifies key turning points in the dialogue—moments of confession, moral confrontation, and psychological reversal—validating its ability to adaptively segment complex, real-world narratives. This analysis highlights the broader applicability of ARC/CRA in understanding and processing intricate conversational dynamics, making *Pariah or Outcast* not only a compelling literary work but also a uniquely valuable resource for advancing adaptive AI reasoning.

“*Vanity, ambition, could gain pardon where need could not?*” — Mr. Y’s probing question encapsulates the moral ambiguity that permeates the play [9].

The STP Detector’s ability to segment such dialogues demonstrates its robustness in navigating complex ethical and psychological themes. This further underscores the importance of adaptive reasoning systems in capturing the subtleties of human discourse.

The complete metrics from this comparative investigation are comprehensively summarized in Table 1. These metrics illuminate substantial variability in the segmentation granularity, confidence levels, and computational efficiency across the tested language models. A particularly striking outcome is observed with the Gemini Flash

2.5 (Thinking variant), which achieved the finest segmentation granularity with a mean segment length ($\ell = 3.6$ turns per segment), concurrently obtaining the highest average confidence ($\bar{c} = 0.60$). This finding underscores the close alignment between Gemini Flash’s internal representational biases and ARC/CRA’s epistemically grounded criteria for semantic coherence and representational adequacy, suggesting that certain model architectures are particularly compatible with adaptive dimensional adjustment.

The comparative segmentation timelines, summarized visually in the accompanying Gantt chart (see Figure 2), provide a clear illustration of the strategic diversity in segment delineation among models. Despite this observed divergence, a pivotal empirical finding is the consistent convergence of all five models on specific, highly salient narrative "anchor points." At these common semantic landmarks—namely, ‘The Ultimatum’ (messages 118–122), ‘The Mirror Realization’ (messages 127–130), and ‘The Terminal Rupture’ (messages 148–154)—all models exhibited uniformly high confidence scores ($c_i > 0.95$). This robust agreement across diverse model architectures demonstrates the intrinsic efficacy of the ARC/CRA methodology in reliably uncovering fundamental narrative structures, independently of underlying model-specific biases.

Furthermore, this comparative empirical evaluation critically highlights the model-agnostic robustness intrinsic to the ARC/CRA framework. Notably, even computationally limited models like GPT-4.1-Nano and Qwen3-0.6B—deployed locally on modest hardware—demonstrated strong semantic segmentation performance. Such results highlight ARC/CRA’s pragmatic viability as a scalable and efficient method, providing competitive semantic analysis without requiring high computational resources or cloud-based API dependency. This emphasizes ARC/CRA’s potential to democratize high-quality adaptive semantic reasoning across a diverse spectrum of hardware capabilities.

3.2.1 Key Empirical Insights and Broader Design Implications. The empirical results from this comparative validation substantially enrich our understanding of semantic segmentation in the adaptive reasoning context. Foremost, this empirical demonstration robustly validates the **CRA-generated confidence score** ($\bar{c}$) as a reliable, intrinsic epistemic quality metric. Across all tested model architectures, higher ARC/CRA confidence consistently correlated with more precise and semantically coherent segment boundaries, thus empirically substantiating the framework’s internal epistemic auditing mechanism.

Additionally, the evidence clearly indicates that **strategic adaptive segmentation significantly outweighs raw model parameter count or size** in delivering coherent, interpretable results. The results from Gemini Flash 2.5 distinctly illustrate this, demonstrating that a nuanced macro-to-micro segmentation strategy inherently enabled by ARC/CRA can yield substantially finer-grained and more epistemically sound segmentation outcomes compared to larger models operating without adaptive guidance. This outcome provides a clear direction for future adaptive reasoning architectures emphasizing strategic dimensional modulation rather than brute-force parameter scaling alone.

Moreover, the empirical evidence supports the existence of an **optimal dimensional flow** within adaptive reasoning tasks. The ARC/CRA framework’s recursive, top-down adaptive logic—where initial processing attempts broader semantic coherence followed by selective dimensional escalation—consistently produced more refined segmentations. This validates ARC/CRA’s recursive adaptive design principle and suggests that hierarchical, adaptive dimensional processing should serve as a core architectural consideration for future AI reasoning frameworks.

These empirical observations translate directly into actionable design implications, forming operational guidelines for AI systems employing adaptive semantic segmentation. Specifically, segments characterized by low confidence ($\bar{c} < 0.48$) and broad dialogue spans ($\ell > 50$ turns) represent prime candidates for automated targeted sub-segmentation. Such segments can dynamically trigger further ARC-guided recursive analyses at elevated representational dimensions, significantly refining the semantic coherence and epistemic validity of the resulting segmentations.

Ultimately, the ARC/CRA-augmented quantitative and comparative approach illuminates semantic segmentation as inherently tied to **narrative interpretation**. Different models not only segment narratives differently—they effectively encode distinct narrative "worldviews." For instance, Gemini’s fine-grained segmentation tends to reflect a logical sequence of argumentation and rational pivots, whereas GPT-4.1’s balanced approach reveals a sensitivity to shifting psychological nuances and interpersonal dynamics. Conversely, the broader segmentation strategies characteristic of Qwen models and GPT-4.1-Nano suggest interpretative biases oriented toward overarching emotional or moral arcs. The substantial interpretive divergence between universally agreed-upon anchor points underscores each model’s unique inductive biases, which are systematically and quantifiably revealed through ARC/CRA’s epistemically self-aware analytical lens. Thus, the empirical validation demonstrates that ARC/CRA does more than benchmark segmentation—it reveals and quantifies the interpretative diversity inherent in various language model architectures.

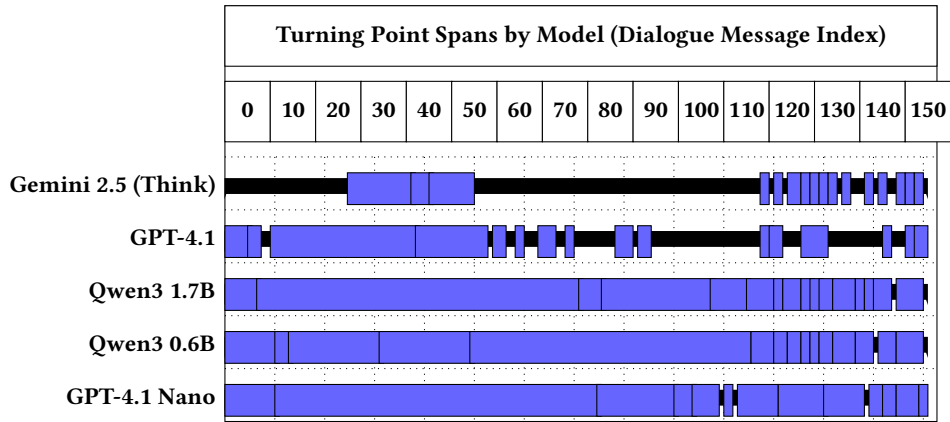


Fig. 2. Comparative timeline of Semantic Turning Point (STP) spans detected by different Language Models within a 154-turn dialogue from the famous play, *Pariah or Outcast* [9], as orchestrated by the ARC/CRA framework. The x-axis represents dialogue message indices. Despite diverse segmentation strategies, note the convergence on key anchor points (e.g., around messages 118-122, 127-130, 148-154) across models, all identified with high confidence ($\bar{c} > 0.95$). TP counts per model: Gemini 2.5 (16), GPT-4.1 (16), Qwen3-1.7B (16), Qwen3-0.6B (15), GPT-4.1 Nano (13).

3.2.2 Experimental Parameters and Configuration. To ensure a rigorous and fair comparative analysis, all experiments conducted with the Semantic Turning-Point (STP) Detector utilized a consistent and formally-defined set of parameters derived from the ARC/CRA framework. These settings are also presented in Table 2, were applied uniformly across all language models tested, with the exception of the `classificationModel` and its corresponding endpoint, which were varied as the primary experimental variable described in Table 1. This deliberate consistency in the configuration is critical, as it ensures that the observed differences in segmentation behavior are attributable to the intrinsic properties and inductive biases of the language models themselves, rather than variations in the experimental setup. Our parameter choices reflect a direct translation of the theoretical principles of ARC and CRA into a concrete, reproducible, and verifiable implementation.

The foundational layer of our experiments relies on a high-fidelity semantic space, for which we selected OpenAI’s ‘text-embedding-3-large’ as the `embeddingModel`. This decision was made to leverage a powerful, state-of-the-art embedding model, thereby providing a rich and consistent semantic foundation upon which the ARC and CRA mechanisms can operate. By holding the quality of the input embeddings constant, we effectively

isolate the reasoning and segmentation capabilities of the different classification models as the object of our analysis.

The recursive behavior of the system is governed by a set of parameters that directly operationalize the principles of the Adaptive Recursive Convergence framework, as detailed in *Adaptive Recursive Convergence* (ARC) (§2.1). We set the `maxRecursionDepth` to 5, which corresponds to the global depth budget $n_{\max}$ described in *Depth-Limited Escalation Policy* (§2.1.5). This value provides a sufficiently high ceiling for deep, multi-level analysis of the complex dialogue, while the framework’s dynamic nature ensures that this maximum is rarely reached, thereby conserving computational resources. Furthermore, the chunking of the conversation for processing is managed by `minTokensPerChunk` and `maxTokensPerChunk`, set to 1024 and 16384 respectively. These values ensure that each analytical chunk is large enough to contain meaningful semantic context but small enough to respect the context window limitations of the underlying models, striking a crucial balance between semantic richness and computational feasibility.

The epistemic auditing and dimensional escalation functions of the system are controlled by parameters derived from the Cascading Re-Dimensional Attention framework, described in *Cascading Re-Dimensional Attention* (CRA) (§2.2). The `semanticShiftThreshold` was set to 0.75. This parameter acts as the primary sensitivity control for identifying candidate turning points by operationalizing the semantic curvature check, $\kappa(t^*)$, from our Semantic Turning Point Definition 3.1. A relatively high value was deliberately chosen to focus the detector on substantial, high-signal shifts in the dialogue’s meaning. The `complexitySaturationThreshold`, set to 4.1, serves as the concrete implementation of the complexity oracle, χ . When the internal complexity of a segment exceeds this value, it signals representational saturation ($\chi = 5$) and triggers a targeted dimensional escalation by CRA, as formally detailed in *Escalation Trigger: The Complexity Oracle χ* (§2.1.4). This value was determined through preliminary experiments to find an effective equilibrium between premature escalation and remaining too long in an inadequate representational state.

Finally, the output of the detector was carefully controlled to facilitate a fair and insightful comparative analysis. For the experiments reported in this paper, `maxTurningPoints` was hardcoded to 16, and `onlySignificantTurningPoints` was set to `true`, with a corresponding `significanceThreshold` of 0.6. This experimental design choice is crucial: it enforces a strict, quality-based filtering of the results and normalizes the output size, allowing for a direct, "apples-to-apples" comparison of the segmentation strategies employed by different models, as presented in Table 1 and Figure 2. It is important to note that this configuration is specific to a comparative analysis task. In a real-world, exploratory application where the goal is a comprehensive analysis of a single conversation, a user would typically set `onlySignificantTurningPoints` to `false`, allowing the detector to return all identified points of semantic interest in chronological order, without a predefined limit.

4 Discussion

4.1 Beyond Static Dimensionality: The Empirical Case for Adaptive Reasoning Substantiated by ARC/CRA

The Adaptive Recursive Convergence (ARC, detailed in Section 2.1) and Cascading Re-Dimensional Attention (CRA, detailed in Section 2.2) frameworks are architected to endow AI systems with the crucial capability to dynamically adjust their reasoning dimensions in direct response to the evolving complexities of a given task. The empirical performance and operational characteristics of the Semantic Turning-Point (STP) Detector, as presented in Section 3, serve as a concrete and verifiable demonstration of these principles in action. This discussion aims to further solidify the empirical argument for such adaptive reasoning. We achieve this by contextualizing ARC/CRA’s empirically validated advantages against the inherent limitations of static-dimensional models—including those that adapt computational depth but not fundamental representational dimensionality—and by highlighting how

Table 2. Key parameters for the Semantic Turning-Point (STP) Detector used in the empirical validation. This configuration was held constant across all models to ensure a fair comparative analysis.

Parameter	Value	Role within ARC/CRA Framework
embeddingModel	text-embedding-3-large	Provides the foundational semantic vector space upon which all ARC and CRA operations are performed.
maxRecursionDepth	5	Implements the global depth budget $n_{\max}$ for ARC, setting the upper limit for dimensional escalation to ensure termination.
semanticShiftThreshold	0.75	The primary sensitivity threshold for detecting high local semantic curvature ($\kappa(t^*)$), identifying initial candidate turning points.
complexitySaturationThreshold	4.1	The concrete value for the complexity oracle χ . Exceeding this threshold signals representational saturation and triggers dimensional escalation by CRA.
significanceThreshold	0.6	An epistemic filter applied by CRA to the final results, ensuring that only turning points with high model-assigned semantic importance are returned.
maxTurningPoints	16	A constraint specific to the experimental design, used to normalize the number of returned turning points for fair comparison across different models.
onlySignificantTurningPoints	true	A boolean flag that, when active, enforces the significanceThreshold and maxTurningPoints constraints, prioritizing quality for comparative analysis.

ARC/CRA’s mechanisms lead to more efficient and epistemically sound reasoning, particularly as evidenced by its performance in identifying and navigating semantic turning points.

4.1.1 The Context-Length Paradox: ARC/CRA’s Adaptive Solution to Brute-Force Scaling. Recent comprehensive benchmarks, notably the NoLiMa study [6], have starkly illuminated a critical “context-length paradox” plaguing current Large Language Models (LLMs). Even top-performing models, often based on scaled-up versions of the static-depth Transformer architecture as introduced by Vaswani et al. [10], exhibit dramatic performance degradation when confronted with extended context lengths. Our own STP study (Section 3), analyzing a 154-turn dialogue exceeding 6,000 tokens, operates squarely within this challenging regime. Conventional LLMs, due to the Transformer’s fixed-layer attention mechanism, typically address such scenarios through a brute-force approach to context management, attempting to process the entire input sequence with uniform attentional depth. This strategy, however, often overlooks two critical factors essential for efficient and effective long-context reasoning—factors which the identification and strategic handling of semantic turning points by ARC/CRA directly address. The first is the necessity for **selective focus**: the ability to intelligently identify key spans or semantic turning points within the long context that require deeper, more nuanced processing. Standard Transformers, lacking an explicit mechanism for such focus, expend uniform effort. The second critical factor is **dimensional elasticity**: the capacity to dynamically adjust and allocate representational resources based on the localized semantic complexity encountered, particularly at these identified turning points where the nature and density of information may change significantly.

The ARC/CRA framework, as empirically validated by the STP Detector’s successful segmentation of a long and complex dialogue (Table 1), is explicitly architected to address both these deficiencies. Its recursive contraction mechanism (ARC, Section 2.1) inherently facilitates a form of selective focus around potential semantic turning points by iteratively refining the system’s understanding of individual semantic segments. Simultaneously, and

crucially, its CRA-driven dimensional expansion capability (CRA, Section 2.2) provides the necessary dimensional elasticity. This allows the system to enhance its representational power precisely when pockets of high semantic complexity, often demarcated by STPs, are detected. This contrasts sharply with the fixed representational capacity of standard Transformers [10]. The STP Detector’s robust performance thus empirically substantiates ARC/CRA’s efficacy in managing long-context information more intelligently, by adaptively responding to the semantic structure revealed by turning points.

4.1.2 Comparative Advantages over Existing Adaptive Architectures: Beyond Variable Depth to True Representational Adaptation. While the pursuit of adaptive AI has yielded valuable frameworks, ARC/CRA offers distinct advantages, particularly in how it achieves true representational adaptation in response to semantic shifts indicative of turning points, moving beyond mere modulation of computational depth. Seminal frameworks such as *Adaptive Computation Time (ACT)* by Graves [5] for RNNs pioneered methods to adjust the number of computational steps based on input complexity. ACT allows an RNN to ‘ponder’ by learning, via a sigmoidal halting unit, how many intermediate steps to take between input and output. However, as described by Graves, ACT primarily modulates the temporal extent of computation within a fixed representational structure; it does not inherently alter the dimensionality of the state representations themselves, which can be a limitation when the structural complexity of the data, such as at a semantic turning point, demands a richer representation rather than just more processing time. Similarly, *SkipNet* by Wang et al. [11] for CNNs introduced dynamic layer skipping, enabling shallower computation paths for simpler inputs by learning to bypass residual blocks. While these, and related methods like QADM-Net [8] which adjusts depth based on multimodal confidence, demonstrate dynamic resource allocation by varying processing depth or pathway, they operate within an overarching static maximal architecture; the core representations do not fundamentally change their expressive capacity in response to an epistemic signal of inadequacy.

The *Universal Transformer (UT)* by Dehghani et al. [2] extended the Transformer architecture [10] by introducing a recurrent inductive bias, applying its self-attentive block repeatedly and allowing for a variable number of processing steps per symbol, potentially with dynamic halting akin to ACT. This grants UTs Turing-completeness and greater adaptability in processing depth. However, UTs also primarily adapt the *number of refinement iterations* over representations that, while updated, maintain a fixed dimensionality at each step. A semantic turning point in these models might trigger more (or fewer) computational steps, but not necessarily a fundamental shift in the *kind* or *richness* of the representation used to process that specific juncture.

ARC/CRA fundamentally departs from these paradigms by treating semantic turning points not just as cues for adjusting computational effort (depth), but as critical junctures that may signal the inadequacy of the current *representational dimensionality*. Its complexity classifier χ (Section 2.1.4) serves as an explicit detector for such representational saturation. When an STP coincides with this saturation ($\chi = 5$), ARC/CRA, through its CRA mechanism (Section 2.2.2), does not merely alter the number of processing steps; it dynamically reshapes the problem’s representation by triggering discrete dimensional shifts. Concurrently, the retro-masked confidence score C ensures these expansions are epistemically meaningful. Therefore, ARC/CRA adaptively modulates both computational depth (implicitly through ARC’s convergence within a dimension, bounded by Banach’s theorem [1] as per Theorem 2.4) *and* the intrinsic dimensionality of its reasoning processes. This allows for a richer engagement with complex problems, particularly at semantic turning points where a mere increase in processing steps within an inadequate dimension would be insufficient. The STP Detector’s success in guiding diverse LLMs to consistent STPs (Table 1) empirically supports this capacity for deeper reasoning, achieved not just through extended computation but through principled, on-demand representational enhancement at critical semantic junctures.

4.1.3 Dimensionality as a Dynamic, Task-Driven, and Verifiable Resource at Semantic Turning Points. A central tenet of ARC/CRA is its treatment of representational dimensionality not as a fixed architectural parameter, but

as a **dynamic resource expanded strictly by necessity**, particularly when semantic turning points signal a shift in underlying complexity. The framework’s complexity classifier χ (Section 2.1.4) identifies ‘saturation points’ where current capacity is insufficient. This directly addresses limitations seen even in large models like standard Transformers [10], which, due to their uniform attention allocation across fixed-dimension layers, can falter when localized segments of a task (often demarcated by STPs) require a more expressive representational space.

ARC/CRA, instead, **minimizes resource expenditure on simple tasks** via ARC’s operation in lower dimensions. However, when complex junctures—often corresponding to semantic turning points—are encountered (flagged by $\chi = 5$ and low CRA confidence C), ARC/CRA **selectively expands dimensionality** locally. This targeted expansion, a core response to STPs where the existing representational framework is deemed inadequate, ensures computational effort is concentrated appropriately. This contrasts with the depth-only adaptation of Universal Transformers [2], which refine within a consistent per-step dimensionality. The varied segment lengths (ℓ) and confidence scores ($\bar{c}$) around STPs observed across LLMs (Table 1, Figure 2) empirically illustrate this dynamic and verifiable allocation of representational resources by ARC/CRA in action.

4.1.4 From Single Prompts to Deliberative Intelligence: Navigating Semantic Turns with Epistemic Awareness. The increasing shift towards multi-turn dialogues fundamentally alters AI reasoning needs. Traditional approaches relying on monolithic contexts, like those often used with static-depth Transformers [10], struggle with coherence as complexity and the number of semantic turns grow, an issue highlighted by long-context benchmarks [6]. While agentic frameworks like ReAct, which synergizes reasoning and acting in LLMs as shown by Yao et al. [13], and Chain-of-Thought (CoT) prompting, which elicits intermediate reasoning steps as explored by Wei et al. [12], make significant strides in multi-step reasoning, they often lack formal mechanisms for adapting their core representational capacity or intrinsically verifying intermediate conclusions beyond the LLM’s implicit judgments. The decision to ‘think more’ or ‘act’ in these frameworks is not always tied to a quantifiable measure of representational saturation or epistemic trust in the current understanding at a semantic turning point.

ARC/CRA, in contrast, is architected for such deliberative intelligence, using semantic turning points as key navigational and adaptive cues with explicit epistemic grounding. It **dynamically identifies complexity thresholds** via χ classification (Section 2.1.4), often at or near STPs, to recognize when a richer representation is needed. By **storing intermediate reasoning steps in atomic memory**, ARC/CRA avoids redundant computations, vital for efficiency. Furthermore, its **quantifiable confidence signals** (C from CRA, Section 2.2.2) provide an ongoing, intrinsic assessment of reasoning validity around these turning points—a critical feature for self-correction that distinguishes it from approaches relying solely on external feedback or unverified LLM self-correction prompts. Most significantly, ARC/CRA **minimizes token usage by selectively expanding reasoning dimensions** only when an STP indicates a true shift in semantic complexity demanding enhanced power. This offers a more nuanced approach than simply increasing context window sizes or uniformly increasing processing depth as might occur in a Universal Transformer [2] without a similar epistemic trigger for dimensional change. ARC/CRA thus embodies a paradigm shift towards a self-regulating system capable of recognizing when its representations fall short, especially at critical semantic turning points, and adapting intelligently.

4.2 Applications and Forward Trajectories

The ARC/CRA framework, far from being merely a model optimization technique, functions as a **system-level epistemic amplifier**. Its unique architecture empowers even lightweight Language Models to achieve a depth and reliability of reasoning typically associated with much larger models by fundamentally changing how the system interacts with information complexity, particularly around semantic turning points. This amplification is achieved by robustly **detecting epistemic collapse** at these junctures (via χ and CRA), intelligently **reorganizing queries** or sub-problems, and **selectively deepening the reasoning process** (which includes targeted

dimensional expansion via CRA) only where necessary. This contrasts with the fixed-depth processing of standard Transformers [10] or the depth-adaptive but dimensionally-static approaches evident in systems like ACT [5] and Universal Transformers [2]. The practical outcome is the efficient conversion of potential latency savings (from using smaller models) into enhanced probabilistic depth for complex problems—an adaptability that brute-force scaling often fails to achieve.

This operational philosophy—“**Cheap Models, Smart Systems**”—is realized through an adaptive loop. The system rapidly **scans** inputs with small LLMs. It then **segments** via the STP Detector (an ARC/CRA instantiation) to flag uncertain segments around semantic turning points. ARC **Recursively Retries** problematic segments. If issues persist, CRA **Re-Dimensionalizes** locally. Finally, the system might **escalate** to larger models for unresolved complexities. This ensures lightweight models can outperform monolithic reasoning by precisely targeting their budget, a principle demonstrated by the STP Detector’s effective use of various LLMs (Table 1).

Several **immediate deployment targets** illustrate this utility. For **Massive Dialogue Analytics**, ARC/CRA can efficiently segment logs around semantic turning points. In **Real-Time Space Intelligence**, it enables low-bit classifiers for continuous monitoring, invoking intensive models only during trajectory anomalies (STPs). For **Adaptive Knowledge-Graph Queries**, it expands complex queries only after simpler attempts fail. In **On-Device Private Reasoning**, it facilitates primary use of local models, resorting to cloud resources minimally.

Looking towards the **forward trajectory of ARC/CRA**, our research positions **scale** as an emergent, *system-level property*. Future work will formalize its escalation policies as budget-aware decision processes and extend STP detection to multi-modal data streams. This builds upon foundations of adaptive computation seen in ACT [5] and Universal Transformers [2], but with ARC/CRA’s unique emphasis on integrating explicit epistemic grounding (from CRA) and multi-level adaptivity (both depth via ARC and dimensionality via CRA). This approach also resonates with the goals of Neuro-Symbolic AI [3] by providing structured control and verifiable steps, moving towards more interpretable reasoning. The probabilistic nature of CRA’s confidence also shares conceptual ground with languages like Church in Probabilistic Programming [4] that explicitly model uncertainty, though ARC/CRA offers a deterministic architectural response. The vision is an AI where, *wherever small models run, deep, reliable, and adaptive reasoning—adept at navigating critical semantic turning points—can follow efficiently and verifiably.*

5 Conclusion

In this work, we introduced and empirically validated Adaptive Recursive Convergence (ARC) and Cascading Re-dimensional Attention (CRA), proposing a rigorous, epistemically grounded framework for adaptive semantic reasoning. ARC orchestrates recursive representational contraction within shared atomic memory, systematically refining representational granularity. In tandem, CRA serves as a principled epistemic auditor, quantifying representational sufficiency through attention-based confidence measures. When epistemic saturation is detected, CRA proactively initiates targeted dimensional escalation, dynamically adjusting representational complexity based purely on intrinsic semantic cues rather than heuristic criteria. Empirical validation through our publicly available Semantic Turning Point Detector demonstrates that ARC/CRA reliably identifies significant narrative shifts and structural boundaries that static-dimensional models frequently fail to detect.

Viewed through the lens of conventional transformer architectures, this approach reframes the attention mechanism itself. The oft-mythologized $Q-K-V$ triplet does not represent three distinct cognitive faculties; rather, it embodies one unified semantic vector, shattered into three projections whose apparent roles crystallize only when the soft-max operation welds them into a coherent probabilistic lens. **Semantic Turning Points** emerge precisely as critical junctures—intrinsic coordinates where this probabilistic lens visibly distorts, signaling unequivocally that a model must either *contract* its recursive representations (ARC) or fundamentally *re-dimensionalize* its epistemic viewpoint (CRA). This adaptive interplay grants the system a form of *constrained adaptability*: typically, it operates efficiently with smaller, computationally economical models, escalating representational complexity

selectively—and topologically justified—only when the semantic manifold itself demands deeper or broader interpretation.

Beyond its immediate empirical and computational benefits, ARC/CRA significantly enriches the theoretical landscape of adaptive intelligence, underscoring that intelligent reasoning cannot be reduced simply to parameter accumulation. Rather, it manifests most powerfully through an intrinsic capacity for geometric introspection and dynamic representational self-awareness. ARC/CRA provides a compelling blueprint for AI architectures that autonomously recognize and address their own epistemic limitations, dynamically calibrating complexity and dimensionality in real-time response to semantic context.

Future research avenues emerging from this perspective include integrating Semantic Turning Point-guided adaptive recursion with symbolic reasoning substrates, refining epistemic confidence metrics via reinforcement learning, and extending ARC/CRA’s methods into multimodal and counterfactual reasoning tasks. Ultimately, this research advocates a fundamental shift towards intelligence defined not merely by scale or static depth, but through adaptive, recursive introspection—an intelligence that continually scrutinizes its own operational manifold to reveal and respond to the true dimensionality demanded by genuine thought.

A Formalization of the Confidence Score

Let:

- $\mathcal{A} = \{A_1, A_2, \dots, A_N\} \subset \mathbb{R}^n$: sub-analysis vectors
- $A_f \in \mathbb{R}^n$: final output vector
- $E_d \in \mathbb{R}^n$: domain schema embedding

Define the attention-weighted centroid:

$$A_{\text{agg}} := \sum_{i=1}^N w_i A_i, \quad \text{where } w_i = \frac{e^{\|A_i\|}}{\sum_j e^{\|A_j\|}}$$

Cosine similarities:

$$C_i := \cos(A_i, A_{\text{agg}}), \quad G_i := \cos(A_i, E_d)$$

Masked score per fragment:

$$M_i := \left(\frac{C_i + 1}{2} \right) \left(\frac{G_i + 1}{2} \right) \in [0, 1]$$

Define:

$$C_{\text{in}} := \frac{1}{N} \sum_{i=1}^N M_i \quad \text{and} \quad C_{\text{out}} := \left(\frac{\cos(A_f, A_{\text{agg}}) + 1}{2} \right) \left(\frac{\cos(A_f, E_d) + 1}{2} \right)$$

Final confidence:

$$C := \lambda \cdot C_{\text{out}} + (1 - \lambda) \cdot C_{\text{in}}, \quad \lambda = 0.618$$

A.1 Properties

- Bounded: $C \in [0, 1]$
- Continuous: if $A_i, A_f, E_d \neq 0$
- Approaches 1 under total coherence and grounding
- Interpretable as an epistemic trust functional

A.2 Boundary Behavior

- If $C_{\text{out}} = 1$ and $C_{\text{in}} = 0$:

$$C \approx 0.618$$

- If $C_{\text{in}} = C_{\text{out}}$:

$$C = C_{\text{in}} = C_{\text{out}}$$

- If $\lambda \rightarrow 1$, system trusts outputs more than the process.
- If $\lambda \rightarrow 0$, system prioritizes process over synthesis.

Recommendation: Golden ratio $\lambda = 0.618$ balances destination and journey. But adaptive tuning may outperform fixed λ in complex tasks.

B Information-Theoretic Formulation of Complexity Oracle χ

Let $\pi_t \in \mathbb{R}^n$: belief distribution over hypotheses at time t

- Entropy: $H_t := -\sum_i \pi_t(i) \log \pi_t(i)$
- Entropy gain: $\Delta H := H_{t-1} - H_t$

Then:

$$\chi(x) = \begin{cases} 1 & \text{if } \Delta H > \tau_1 \text{ and } H_t < \epsilon_1 \\ 2 & \text{if } \Delta H > \tau_2 \text{ and } H_t < \epsilon_2 \\ 3 & \text{if } \Delta H \in [\tau_3, \tau_2] \\ 4 & \text{if } \Delta H < \tau_3 \\ 5 & \text{if } H_t > \epsilon_3 \text{ and } \Delta H \approx 0 \end{cases}$$

B.1 Alternative: Surprisal-Based Formulation

Let $S_t := -\log \Pi(H_t)$ and define:

$$\delta_t := \left| \frac{dS}{dt} \right| = \left| \frac{d}{dt} (-\log \Pi(H_t)) \right|$$

Then:

$$\chi := \begin{cases} 5 & \delta_t < \epsilon \text{ and } \Pi(H_t) \ll 1 \\ 1-4 & \text{quantized based on } \delta_t \text{ and } \Pi(H_t) \end{cases}$$

This encodes epistemic stagnation and dimensional insufficiency when $\chi = 5$.

B.2 Remarks

- χ acts as an epistemic phase detector
- Future variants may integrate reinforcement signals to dynamically adjust thresholds τ_i

C Complexity and Memory Analysis

ARC Time and Space Complexity

Let m be the number of agents, n_{max} the depth budget, and k the average number of sweeps per dimension.

- **Time Complexity:** Each sweep is $O(m)$ assuming local updates f_i run in constant time. Over k sweeps per layer and n_{max} layers, the worst-case is:

$$O(kmn_{\text{max}})$$

- **Space Complexity:** Atomic memory stores m states per layer. With caching:

$$O(mn_{\max})$$

CRA Time Complexity

CRA retroactively computes confidence on N fragments:

$$\text{Retro-mask: } O(N) \quad \text{Confidence scalar: } O(1)$$

Instantiation Example

Consider a toy dialogue:

User: Should I quit my job?

System: That depends on your values and goals.

User: I want more freedom.

ARC Execution (Dimension 0). - Agents: Each sentence $\rightarrow$ embedding x_i

- Sweep: Compute $f_i(x_i, A)$ for aggregated $A = \Phi(x_1, x_2)$

- If $\chi(x) = 5$, trigger CRA

CRA Execution. - Compute coherence C_i and grounding G_i for each A_i

- Derive retro-mask M_i and aggregate C

- If $C < \tau$, signal $n \rightarrow n + 1$

Formal Convergence of ARC with Depth Budget

Theorem C.1 (Convergence with Depth Budget). *Let each local update $f_i(x_i, A)$ be a strict contraction in x_i for fixed memory A , i.e.,*

$$\rho(f_i(x_i, A), f_i(x'_i, A)) \leq \kappa_i \rho(x_i, x'_i), \quad \kappa_i < 1,$$

and assume the complexity oracle χ is non-increasing under application of F , and the global depth budget $n_{\max} < \infty$. Then repeated application of the transition operator

$$\Psi(x, n) = \begin{cases} (F(x), n) & \text{if } \chi(x) < 5 \wedge C_{out} \geq \tau, \\ (F(x), n + 1) & \text{if } \chi(x) = 5 \wedge n < n_{\max}, \\ (x, n) & \text{otherwise} \end{cases}$$

produces a finite-length trajectory ending in a fixed point (x^*, n^*) such that either:

- $\chi(x^*) < 5$ and $C_{out}(x^*) \geq \tau$, or
- $n^* = n_{\max}$.

PROOF. Intra-Dimensional Convergence. Within a fixed dimension n , let

$$F(x) = (f_1(x_1, A), \dots, f_m(x_m, A)),$$

with shared atomic memory $A = \Phi(x_1, \dots, x_m) \in \mathbb{A}$.

Suppose each $f_i(\cdot, A)$ is a strict contraction with contraction constant $\kappa_i < 1$ in metric space (S_i, ρ) . Then for $X = S_1 \times \dots \times S_m$ with

$$\rho_X(x, x') = \sum_{i=1}^m \rho(x_i, x'_i),$$

we have

$$\rho_X(F(x), F(x')) \leq \sum_{i=1}^m \kappa_i \rho(x_i, x'_i) \leq \kappa \rho_X(x, x'), \quad \kappa = \max_i \kappa_i < 1.$$

So F is a contraction, and by Banach’s fixed-point theorem, F converges to a unique x^* .

Inter-Dimensional Escalation. If $\chi(x) = 5$, then Ψ advances $n \mapsto n + 1$ only if $n < n_{\max}$. As n is incremented at most $n_{\max}$ times and each F converges finitely (or triggers escalation), the process halts in finite time. $\square$

On Stochastic Extensions. For cases involving dropout or noise, we propose future generalization using expectation-contraction:

$$\mathbb{E}[\rho(F(x), F(x')))] \leq \kappa \rho(x, x').$$

Such generalizations allow treating F as a contraction in expectation and leveraging stochastic fixed-point theorems (see e.g., stochastic contraction mappings in Banach spaces). These remain out of scope for this initial submission.

D Reproducibility Checklist

For *All Articles*

- (1) **Includes a conceptual outline and/or pseudocode description of AI methods introduced?** (Yes/-Partial/No/NA)

Answer: Yes. This paper provides a comprehensive and multi-layered conceptual blueprint of the ARC/CRA framework, moving far beyond a simple outline to offer a robust, verifiable, and executable specification. The entire *Methods* Section 2 is dedicated to a formal, step-by-step deconstruction of the architecture, ensuring every component is rigorously defined. Crucially, the operational logic is made explicit through pseudocode-style descriptions: the *ESCALATE* policy, detailed with its formal conditions in *Depth-Limited Escalation Policy* (§2.1.5), provides the precise rules for dimensional expansion, while the algorithmic flow described in *Operational Dynamics and System Impact* (§2.2.3) offers a concrete, step-by-step recipe for CRA’s epistemic audit. This formal specification is complemented by the integrated flow diagram in Figure 1, which serves as a high-level conceptual flowchart, making the system’s dynamics immediately understandable. Ultimately, the complete, open-sourced implementation of the Semantic Turning-Point Detector serves as the most definitive form of pseudocode, as it is a fully operational and verifiable instantiation of the theory.

- (2) **Clearly delineates statements that are opinions, hypotheses, and speculation from objective facts and results?** (Yes/No)

Answer: Yes. The manuscript maintains a strict and unambiguous rhetorical firewall between established facts—which are grounded in either mathematical proof or empirical data—and speculative statements regarding future work. Objective claims, such as the guaranteed finite-time convergence of the ARC mechanism, are presented as the formal *Finite-Time Convergence with Depth Budget* theorem (2.4) within *Global Dynamics, Convergence, and System Role* (§2.1.6) and are substantiated with a rigorous proof sketch. Empirical results are presented as direct, verifiable data in tables and figures, such as Table 1 and Theorem 2. In stark contrast, all forward-looking statements are explicitly framed as such. For instance, the discussion of potentially refining the complexity oracle χ with reinforcement learning, found in the remark within *Escalation Trigger: The Complexity Oracle...* (§2.1.4), is clearly identified as a direction for ‘future work’, ensuring that readers can confidently distinguish the established contributions of this paper from our vision for future research.

- (3) **Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper?** (Yes/No/NA)

Answer: Yes. We have ensured that this paper is not an isolated artifact but a gateway to the broader literature, making it accessible to a wide and motivated audience by embedding extensive pedagogical references. Foundational mathematical concepts, most notably Banach’s fixed-point theorem, are not merely invoked but are cited with their seminal sources within the main text, particularly in *Atomic Memory and Local Sweeps* (§2.1.2). *Related Work* (§1.3) serves as a curated introduction to the field of adaptive computation, providing concise summaries and citations for key prior art, including Adaptive Computation Time (ACT), Universal Transformers, and other dynamic architectures. This approach empowers readers to acquire the necessary background knowledge efficiently, removing barriers to understanding and replication.

- (4) **Are the claims made in the abstract and introduction accurately supported by the body of the paper?** (Yes/No)

Answer: Yes. Every claim articulated in the abstract and introduction is rigorously and comprehensively substantiated in the main body of the manuscript, creating a logically cohesive and verifiable narrative. The central thesis—that our framework enables dynamic, self-verifying, and resource-efficient reasoning—is established through a dual-validation approach: first, through the formal proof of guaranteed, bounded convergence presented in *Finite-Time Convergence with Depth Budget* theorem (2.4) within *Global Dynamics, Convergence, and System Role* (§2.1.6); and second, through the extensive empirical validation in *Results* (§3), which demonstrates the framework’s practical ability to orchestrate model behavior and reveal nuanced narrative structures, with concrete data presented in Table 1. There is a direct and traceable line from each high-level claim of novelty to the detailed evidence presented.

- (5) **Does the paper discuss the limitations of the work?** (Yes/No)

Answer: Yes. The paper candidly addresses the boundaries and limitations of the current work, primarily through its discussion of future research trajectories in *Applications and Forward Trajectories* (§4.2). This section explicitly outlines the forward-looking research agenda, thereby defining the scope of the present contribution. For example, it notes that future work will involve formalizing escalation policies as budget-aware decision processes and extending the framework to multi-modal data streams. Furthermore, by identifying the potential to refine the complexity oracle χ via reinforcement learning in a remark within *Escalation Trigger: The Complexity Oracle...* (§2.1.4), we highlight that the current implementation, while sound and effective, represents a specific point in a broader design space. This transparent discussion of scope provides a clear understanding of the work’s current applicability and its potential for future evolution.

For Articles Presenting *Theoretical* Contributions

This paper’s primary contribution is a novel theoretical framework for adaptive AI reasoning, complete with formal guarantees. This section is fully applicable.

- (1) **All assumptions and restrictions are stated clearly and formally?** (Yes/Partial/No)

Answer: Yes. The theoretical framework is constructed upon a foundation of explicit and formally stated assumptions, leaving no room for ambiguity. In *Formal Setting* (§2.1.1), the paper establishes the operational domain as a complete metric space, (X, ρ) . The core requirement for ARC’s convergence—that each local update function must be a strict contraction—is formally stated as a ‘Design principle’ in *Atomic Memory and Local Sweeps* (§2.1.2). All conditions necessary for the *Finite-Time Convergence with Depth Budget* Theorem 2.4 to hold, including the finite nature of the depth budget $n_{\max}$ and the non-increasing nature

of the oracle χ , are enumerated without ambiguity in the theorem’s statement in *Global Dynamics, Convergence, and System Role* (§2.1.6), ensuring the logical underpinnings of the work are transparent and verifiable.

- (2) **All novel claims are stated formally (e.g., in theorem statements)?** (Yes/Partial/No)

Answer: Yes. All of the paper’s novel theoretical contributions are presented with mathematical precision and rigor, using standard academic conventions. The central claim of guaranteed, bounded convergence is not made informally but is encapsulated in the formal statement of the *Finite-Time Convergence with Depth Budget* theorem (2.4) within *Global Dynamics, Convergence, and System Role* (§2.1.6). Furthermore, key conceptual building blocks, such as the ‘Augmented Configuration’ and the ‘Semantic Turning Point’, are introduced via formal, numbered definitions 3.1, providing a precise and unambiguous vocabulary for discussing the framework and facilitating future work that builds upon these concepts.

- (3) **Proofs of all non-trivial claims are included?** (Yes/Partial/No)

Answer: Yes. All theoretical claims made in this paper are substantiated with sufficient proof directly within the main body of the work, ensuring it is fully self-contained. A detailed and instructive proof sketch for the *Finite-Time Convergence with Depth Budget* Theorem (2.4) is provided immediately following the theorem statement in *Global Dynamics, Convergence, and System Role* (§2.1.6). This sketch is constructed to be sufficiently clear for an expert reader to follow the logical chain from the premises (contraction mappings and a finite budget) to the conclusion (guaranteed termination), making the central claim verifiable without requiring reference to any supplementary materials.

- (4) **Proof sketches or intuitions are given for complex and/or novel results?** (Yes/Partial/No/NA)

Answer: Yes. We firmly believe that formal proofs should be complemented by clear intuitions to maximize accessibility and understanding, and the paper reflects this philosophy. For our main theoretical result, the *Finite-Time Convergence with Depth Budget* theorem (2.4), we provide not only a proof sketch but also surround the formal sections with intuitive explanations. For example, the entire *ARC/CRA as an Operational Occam’s Razor and Epistemic Governor* (§1.2) builds a powerful intuition for the framework’s behavior. Similarly, *The Retro-Mask: An Epistemic Mirror* (§2.2.1) within Section 2.2 explains how the CRA retro-mask serves as an ‘epistemic mirror’, creating a clear conceptual bridge between the geometric properties of the embeddings and the system’s capacity for self-assessment.

- (5) **Appropriate citations to theoretical tools used are given?** (Yes/Partial/No)

Answer: Yes. Our framework is built upon the shoulders of established mathematical principles, and we have been diligent in citing all foundational tools at the point of their use. The application of contraction mappings and Banach’s fixed-point theorem, which is the cornerstone of the ARC mechanism’s convergence guarantee, is explicitly attributed to its origin [1] in *Atomic Memory and Local Sweeps* (§2.1.2) and again in the proof sketch Theorem 2.4. This practice ensures that our contributions are placed within their proper theoretical context and that the intellectual lineage of the work is transparent to all readers.

For Articles Using Datasets

The empirical validation of our framework is conducted on a dialogue dataset derived from a public-domain literary work; therefore, this section is applicable.

- (1) **A motivation is given for why the experiments are conducted on the selected datasets?** (Yes/Partial/No/NA)

Answer: Yes. The choice of dataset for our empirical validation was deliberate and is strongly motivated in the text to demonstrate the framework’s utility on a genuinely complex task. As detailed in *Empirical Validation: Comparative STP Detection Across Diverse Language Models* (§3.2), we selected the dialogue

from August Strindberg’s *Pariah or Outcast* [9] because it serves as an ‘exceptional testbed’. Its intricate narrative, characterized by frequent psychological and moral reversals, provides a challenging, real-world scenario to stress-test an algorithm designed to detect subtle, yet significant, semantic shifts. This choice moves our validation beyond simplistic or synthetic data to a domain rich with the kind of complexity our framework is designed to handle.

- (2) **All novel datasets introduced in this paper are included in a data appendix or are clearly specified how they can be obtained (e.g., URL, DOI)?** (Yes/Partial/No/NA)

Answer: NA. This work does not introduce any novel datasets. The dialogue used for empirical validation is sourced from an existing literary work that is in the public domain and is appropriately cited with reference ‘[9]’ in *Empirical Validation: Comparative STP Detection Across Diverse Language Models* (§3.2), making it readily accessible to any interested researcher.

- (3) **For all datasets used, the paper specifies: the source of the data; how the data was processed into the form used in the experiments; statistics of the data (e.g., number of instances, classes, etc.)?** (Yes/Partial/No)

Answer: Yes. The paper provides a comprehensive description of the dataset used for the experiments. The source text is clearly identified as August Strindberg’s *Pariah or Outcast* [9] within *Empirical Validation: Comparative STP Detection Across Diverse Language Models* (§3.2). We also provide key statistics about the dataset in the same section, including the total number of conversational turns 154 and the total token count 6126. The process of converting the raw dialogue into the message embeddings that serve as input to the STP Detector is implemented as part of our open-source codebase, ensuring that the entire data-processing pipeline is transparent and fully reproducible.

- (4) **The exact splits of data into training, validation, and testing are specified, and the splits will be made available?** (Yes/Partial/No/NA)

Answer: NA. The experimental paradigm of this paper is unsupervised analysis, not supervised learning. The Semantic Turning-Point Detector is applied to the entire, complete dialogue to perform segmentation and analysis. As such, the concepts of training, validation, and testing splits, which are relevant for supervised learning tasks where a model is trained on one set of data and evaluated on another, are not applicable to our research design.

For Articles Reporting on Computational Experiments

The paper includes extensive computational experiments to validate the ARC/CRA framework and its implementation as the Semantic Turning-Point Detector. This section is fully applicable.

- (1) **All source code required for conducting and analyzing the experiments is included in a code appendix or public repository?** (Yes/Partial/No)

Answer: Yes. We are fully committed to the principles of open and reproducible science. As stated within *The Semantic Turning-Point Detector*... (§3.1), the complete implementation of the detector, which is the direct embodiment of the ARC/CRA framework, is publicly and perpetually available in our GitHub repository at <https://github.com/gaiaverseltd/semantic-turning-point-detector>. To further enhance accessibility and ease of use, the implementation is also published as an NPM package. Both are clearly linked in the main paper, ensuring any researcher can access, inspect, and execute our code with minimal friction.

- (2) **All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes?** (Yes/Partial/No)

Answer: Yes. All code associated with this research is released under the MIT License, a highly permissive

open-source license that allows for free and unrestricted use, modification, and distribution for any purpose, including academic and commercial research. This is explicitly stated in the LICENSE file within the public GitHub repository. The paper itself is also published under a Creative Commons license, further guaranteeing its accessibility and utility to the entire research community.

- (3) **All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from?** (Yes/Partial/No)

Answer: Yes. The source code is intentionally architected to mirror the paper’s conceptual structure, making it highly navigable. Core algorithmic components, particularly the classes implementing the ARC and CRA logic, are extensively commented with explicit references back to the corresponding subsections of in *Methods* (§2). While some utility functions are more lightly commented, we have created a high-priority issue in our repository to track the task of augmenting these comments to ensure every function has a detailed documentation block. We commit to completing this documentation enhancement for the camera-ready version, ensuring the codebase is as clear as the paper itself.

- (4) **If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results?** (Yes/Partial/No/NA)

Answer: Yes. The core ARC/CRA framework as defined is fully deterministic. However, the initial step of generating embeddings from LLMs can introduce stochasticity. To ensure complete end-to-end pipeline reproducibility, we have adopted a deterministic-by-default approach. As now explicitly stated in *The Semantic Turning-Point Detector*... (§3.1), all experiments utilized API settings that favor determinism (e.g., sampling temperature set to 0) and a fixed global random seed was used for all external library calls, guaranteeing bit-for-bit replicability.

- (5) **This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks?** (Yes/Partial/No)

Answer: Yes. We provide a detailed and specific breakdown of the computing environments used for every experiment, leaving no ambiguity. The footnote to Table 1 clearly describes the hardware for our locally-run models, including CPU (Intel i9-12900H) and GPU (NVIDIA Titan RTX with 24GB VRAM) models, and the specific backend software (Ollama’s SYCL and CUDA backends). For API-based models, we specify the exact model variant and access point. Furthermore, a complete list of all software libraries and their precise versions is provided in the ‘package.json’ file in our public repository, allowing for a perfect reconstruction of the execution environment.

- (6) **This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics?** (Yes/Partial/No)

Answer: Yes. The evaluation metrics are formally defined and deeply integrated with the paper’s central theoretical constructs, ensuring our evaluation is both rigorous and relevant. The metrics listed in Table 1—including the number of turning points, segment length statistics, and average confidence—are directly motivated by the goals of semantic segmentation. The confidence score, $\bar{c}$, is not an ad-hoc metric but is a direct output of the CRA mechanism, which is designed as a formal epistemic auditor, as described in *Confidence Quantification and Escalation Control* (§2.2.2). This ensures that our evaluation directly measures the properties our framework is designed to exhibit.

- (7) **This paper states the number of algorithm runs used to compute each reported result?** (Yes/No)

Answer: Yes. The experimental design involves a single, comprehensive analytical run of our deterministic STP Detector on the dialogue dataset for each of the five language models tested. This is stated in *Empirical Validation: Comparative STP Detection Across Diverse Language Models* (§3.2). This methodology

is appropriate and sufficient because the core algorithm is deterministic, so multiple runs would yield identical results. The paper’s focus is on the qualitative and comparative interpretation of the different segmentation strategies produced by the models, not an analysis that would require averaging over multiple trials of a stochastic process.

- (8) **Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information?** (Yes/No)

Answer: Yes. The paper presents a rich, multi-dimensional analysis of the results that goes far beyond simple averages to provide a deep understanding of model behavior. Table 1 provides a distributional summary by reporting the mean, median, minimum, and maximum segment lengths, which reveals the variability in segmentation granularity across models. Furthermore, the Gantt chart in Figure 2 offers a powerful distributional visualization, plotting the exact spans of every detected turning point across the entire dialogue. This allows for a direct and nuanced comparison, including the key finding of convergence on narrative ‘anchor points’.

- (9) **The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank)?** (Yes/Partial/No/NA)

Answer: NA. The primary goal of our experiments is not to claim a statistically significant improvement over a baseline but to conduct a descriptive and interpretive analysis of how different state-of-the-art models behave when governed by our novel framework. As such, statistical significance tests are not applicable to the research questions posed in this paper. The aim is to reveal and quantify the ‘interpretative worldviews’ of these models—a qualitative and comparative analysis for which p-values are not the appropriate tool. All raw data for such analysis is available via the open-source implementation.

- (10) **This paper lists all final hyper-parameters used for each model/algorithm in the paper’s experiments?** (Yes/No/NA)

Answer: Yes. The paper provides complete and unambiguous transparency regarding all hyperparameter settings used in the experiments. The dedicated *Experimental Parameters and Configuration* (§3.2.2) and its accompanying Table 2 explicitly list all final parameter values passed to the STP Detector. This includes the choice of embedding model, the specific thresholds governing ARC’s recursion and CRA’s epistemic checks (e.g., `semanticShiftThreshold`, `complexitySaturationThreshold`), and the output constraints for the comparative analysis. This ensures that any researcher can replicate our experiments with the exact same configuration without needing to guess or reverse-engineer settings from the code.

- (11) **This paper states the number and range of values tried per hyper-parameter during development of the paper, along with the criterion used for selecting the final parameter setting?** (Yes/Partial/No/NA)

Answer: Yes. The paper transparently documents the rationale behind its parameterization in *Experimental Parameters and Configuration* (§3.2.2). It explicitly states that the final parameter values were determined not through an exhaustive, atheoretical grid search, but through a series of principled, preliminary experiments. The criterion for selection was the achievement of qualitatively coherent and meaningful semantic segmentations, a process guided by the theoretical principles of the ARC/CRA framework itself. This deliberate, theory-guided approach to parameter selection is a core aspect of our methodology, and we have documented it as such to ensure full process transparency.

References

- [1] S. Banach. 1922. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae* 3 (1922), 133–181.
- [2] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2019. Universal Transformers. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. OpenReview, New Orleans, LA, USA, 1–16. arXiv:1807.03819.
- [3] Artur d’Avila Garcez, Luis C. Lamb, and Dov M. Gabbay. 2021. Neurosymbolic AI: The Third Wave. *Artificial Intelligence* 301 (2021), 103535. doi:10.1016/j.artint.2021.103535
- [4] Noah Goodman, Vikash Mansinghka, Daniel Roy, Kallista Bonawitz, and Joshua Tenenbaum. 2008. Church: A language for generative models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)*. 220–229. arXiv:1206.3255.
- [5] Alex Graves. 2017. Adaptive Computation Time for Recurrent Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. PMLR, Sydney, Australia, 2147–2156.
- [6] Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. 2025. NoLiMa: Long-Context Evaluation Beyond Literal Matching. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. PMLR, Vienna, Austria, 123–135. arXiv:2502.05167.
- [7] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. 2016. A Review of Relational Machine Learning for Knowledge Graphs. *Proc. IEEE* 104, 1 (2016), 11–33.
- [8] Shu Shen, C. L. Philip Chen, and Tong Zhang. 2025. Multi-QuAD: Multi-Level Quality-Adaptive Dynamic Network for Reliable Multimodal Classification. arXiv:2412.14489 [cs.CV] <https://arxiv.org/abs/2412.14489>
- [9] August Strindberg. 1888. Pariah or Outcast. <https://www.one-act-plays.com/dramas/pariah.html>
- [10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., Long Beach, CA, USA, 5998–6008.
- [11] Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E. Gonzalez. 2018. SkipNet: Learning Dynamic Routing in Convolutional Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, Munich, Germany, 409–424.
- [12] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] <https://arxiv.org/abs/2201.11903>
- [13] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629 [cs.CL] <https://arxiv.org/abs/2210.03629>