AIR Journal of Mathematics & Computational Sciences, Vol. 2026
Article Unique ID: AIRMCS2026482
DOI: 10.65737/AIRMCS2026482

CHAIN-OF-THOUGHT ICL FOR GEOAI: BREAKING THE HARDNESS BARRIER

Mosab Hawarey
Director, Geospatial Research (*director@geospatial.ch*)

Received: 2026 March 06 | Revised: 2026 March 08 | Approved: 2026 March 08 | Published: 2026 March 09

Abstract

Foundation models for geospatial artificial intelligence (GeoAI) exhibit a striking dichotomy: they excel at pixel-wise tasks such as classification and segmentation, yet struggle with instance-level tasks requiring identification of multiple discrete elements. This limitation, termed the *ICL-Hard barrier*, arises from fundamental computational constraints of constant-depth transformers, which cannot solve problems reducible to sparse parity for more than $J^* = O(\log \log n) \approx 2–4$ objects. The barrier manifests across geospatial domains: geodetic source localization, remote sensing object detection, climate extreme event identification, and multi-modal counting tasks all fail when the number of targets exceeds this threshold.
This paper introduces a unified theoretical framework demonstrating that **chain-of-thought (CoT) prompting** provides a universal mechanism to overcome ICL-Hard barriers across all GeoAI domains. We prove that autoregressive token generation amplifies effective computational depth: generating T intermediate tokens increases effective depth from L to $L + \gamma T$, enabling transformers to escape AC^0/TC^0 circuit limitations and solve previously intractable tasks.
Our main contributions include: (1) the **CoT Depth Amplification Theorem**, proving that T tokens provide effective depth $\Theta(T)$; (2) tight bounds on token complexity, establishing $T = \Omega(J)$ tokens are necessary and $T = O(J \log J)$ sufficient for J-element detection; (3) a smooth success probability scaling law $P(\text{success}) = (1 - e^{-\alpha T/J})^J$; (4) the **CoT Threshold Shift Theorem**, showing the effective threshold increases to $J^*_{CoT}(T) \approx J^* + T$; and (5) equivalence conditions under which CoT matches fine-tuned model performance without parameter updates.
We validate the framework across four GeoAI domains—geodesy, remote sensing, climate science, and multi-modal applications—deriving domain-specific token multipliers $(\kappa \in [2.5, 4.5])$ and practical deployment guidelines. The theory yields eight testable predictions for empirical validation. This work completes a theoretical arc: where previous work established *what* is hard for in-context learning in GeoAI, we now show *how* to overcome these barriers through principled application of chain-of-thought reasoning.

Keywords: chain-of-thought prompting; in-context learning; foundation models; geospatial AI; computational complexity; circuit depth; transformer architectures; remote sensing; climate modelling; multi-modal learning

Citation: Hawarey, M. (2026). Chain-of-Thought ICL for GeoAI: Breaking the Hardness Barrier. AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026482, DOI: 10.65737/AIRMCS2026482.

1. Introduction

1.1 The ICL-Hard Barrier in GeoAI

Foundation models have transformed artificial intelligence, demonstrating remarkable capabilities through *in-context learning (ICL)*—the ability to perform new tasks by conditioning on examples provided in the input context, without updating model parameters. In geospatial artificial intelligence (GeoAI), foundation models such as SatMAE (Cong et al., 2022), Scale-MAE (Reed et al., 2023), Prithvi (Jakubik et al., 2023), Pangu-Weather (Bi et al., 2023), and GeoChat (Kuckreja et al., 2024) have achieved impressive results across diverse Earth observation tasks. Yet a persistent puzzle has emerged: these models excel at certain task categories while systematically failing at others, even when both categories seem superficially similar.

This dichotomy is not accidental. In a series of companion papers (Hawarey 2026a-e), we established a theoretical framework explaining this phenomenon. The framework introduces *Sufficient Statistic Complexity (SSC)* and its attention-computable variant *(AC-SSC)* as measures of task complexity for in-context learning. Tasks fall into two categories:

- **ICL-Easy tasks** possess additive sufficient statistics computable by attention mechanisms. These include spectral classification, semantic segmentation, smooth field prediction, and scene-level captioning. For such tasks, ICL achieves sample complexity $n_{ICL} = \Theta(n_{ERM})$, matching the statistical optimum of empirical risk minimization.
- **ICL-Hard tasks** require combinatorial sufficient statistics that cannot be computed by constant-depth polynomial-size circuits. These include object detection, source localization, extreme event identification, and counting queries. By reduction to sparse parity—a problem known to require exponential circuit size for constant depth (Håstad 1987)—we proved that standard ICL fails for these tasks when the number of discrete elements J exceeds a threshold $J^* = O(\log \log n) \approx 2–4$ for typical context sizes.

The ICL-Hard barrier is universal across GeoAI domains. In geodesy (Hawarey, 2026b), localizing more than 2–3 discrete mass sources from gravity measurements becomes intractable. In remote sensing (Hawarey, 2026c), detecting more than 3–4 objects in satellite imagery exceeds ICL capacity. In climate science (Hawarey, 2026d), identifying more than 3–4 simultaneous extreme events fails systematically. In multi-modal applications (Hawarey, 2026e), counting more than 4–5 objects produces systematic underestimates. This consistency across domains suggests a fundamental computational barrier rather than domain-specific limitations.

1.2 The Open Problem: Can We Break the Barrier?

The theoretical framework in Hawarey (2026a-e) precisely characterizes *what* is hard for in-context learning, but leaves open the question of *how* to overcome these barriers. Hawarey (2026c) explicitly posed this as Open Problem 2:

Open Problem 2 (Chain-of-Thought Analysis): *Prove tight bounds on how chain-of-thought prompting affects ICL complexity. How many intermediate tokens are needed to solve ICL-Hard tasks? Is there a smooth trade-off between token count and success probability?*

Similar observations appeared in Hawarey (2026d-e). Hawarey (2026d) noted that "chain-of-thought prompting may enable sequential event identification, effectively increasing computational depth." Hawarey (2026e) observed that "for counting tasks, CoT decomposition ('Let me count region by region...') may shift the hardness threshold." These hints suggested that chain-of-thought (CoT) prompting might provide a systematic mechanism to overcome ICL-Hard barriers, but no rigorous analysis existed.

Chain-of-thought prompting (Wei et al. 2022) has emerged as a powerful technique for enhancing language model reasoning. By generating intermediate reasoning steps before producing a final answer, models achieve dramatic improvements on arithmetic, commonsense reasoning, and planning tasks. However, the theoretical understanding of *why* CoT works—and specifically how it relates to computational complexity—has remained limited.

1.3 Our Contribution: CoT as Depth Amplification

This paper provides a complete theoretical framework demonstrating that chain-of-thought prompting serves as a *universal mechanism* for overcoming ICL-Hard barriers across all GeoAI domains. Our central insight is that *CoT transforms the computational model: while standard ICL performs a single forward pass of constant depth L, autoregressive token generation creates sequential computation of depth $L + \gamma T$, where T is the number of generated tokens and γ is the depth amplification factor*.

This depth amplification enables transformers to escape the AC^0/TC^0 circuit class limitations that cause ICL-Hard failures. By generating sufficient intermediate tokens, models can cross the $\Omega(\log J)$ depth threshold required for J-element sparse selection tasks.

Our main theoretical contributions are:

Theorem 6.1 (CoT Depth Amplification): We prove that generating T chain-of-thought tokens provides effective computational depth $L_{eff}(T) = L + \gamma T$, where $\gamma \geq 1/(c \log n)$ with $\gamma \to 1$ achievable for well-structured CoT strategies. This formalizes how token generation translates to computational power.

Theorem 6.2 (Token Complexity Lower Bound): We prove that $T = \Omega(J)$ tokens are necessary for J-element detection tasks, establishing fundamental limits on CoT efficiency. This follows from both information-theoretic arguments and computational reductions to sparse parity.

Theorem 6.3 (Success Probability Scaling): We derive the smooth relationship between token count and success probability: $P(\text{all } J \text{ detected} \mid T) = (1 - e^{-\alpha T/J})^J$. This enables principled token budget allocation based on desired accuracy.

Theorem 6.4 (CoT Threshold Shift): We prove that the effective hardness threshold shifts from $J^* \approx 2–4$ to $J^*_{CoT}(T) \approx J^* + T$. With 30 tokens, tasks involving up to 33 elements become tractable; with 100 tokens, over 100 elements can be handled.

Theorem 6.5 (CoT-Fine-tuning Equivalence): We establish conditions under which CoT achieves accuracy equivalent to fine-tuned models: for $J \leq J^*_{CoT}(T)$, CoT matches fine-tuned performance without any parameter updates, at the cost of $T\times$ inference compute.

1.4 Unified Framework Across Domains

A key contribution of this work is demonstrating that CoT provides a *universal* solution across all GeoAI domains identified in (Hawarey, 2026b-e). We systematically apply the framework to:

- **Geodesy (**Hawarey, 2026b**):** Source localization of J discrete mass anomalies from gravity measurements. CoT enables sequential anomaly identification with token multiplier $\kappa_{geo} \approx 3.5$.
- **Remote Sensing (**Hawarey, 2026c**):** Object detection with J objects in satellite imagery. Spatial decomposition strategies achieve token multiplier $\kappa_{RS} \approx 2.5$.
- **Climate Science (**Hawarey, 2026d**):** Detection of J extreme events across spatiotemporal domains. Temporal scanning with compound event analysis requires token multiplier $\kappa_{clim} \approx 4.5$.
- **Multi-Modal GeoAI (**Hawarey, 2026e**):** Counting J objects from natural language queries. Region-by-region enumeration achieves token multiplier $\kappa_{mm} \approx 3.0$.

Despite domain differences, all multipliers fall within a narrow range $\kappa \in [2.5, 4.5]$, reflecting the universal structure of CoT for ICL-Hard tasks: sequential identification of discrete elements through spatial or temporal decomposition.

1.5 Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews preliminaries including the ICL framework from Papers 1-5 (Hawarey, 2026a-e), the ICL-Easy/Hard dichotomy, and relevant circuit complexity background. Section 3 develops the formal theory of chain-of-thought as depth amplification, culminating in Theorem 6.1. Section 4 applies this framework systematically across the four GeoAI domains, deriving domain-specific strategies and token multipliers. Section 5 establishes the tight bounds on token complexity (Theorems 6.2–6.5) and the success probability scaling law. Section 6 synthesizes practical deployment guidelines and eight testable predictions for empirical validation. Section 7 discusses implications, limitations, and connections to the broader research program. Section 8 concludes with a summary and directions for future work.

This paper completes a theoretical arc begun in (Hawarey, 2026a): having established the foundations of ICL characterization, mapped the ICL-Hard landscape across GeoAI domains, we now provide the key to overcoming these barriers. The result is a comprehensive theory of chain-of-thought prompting for geospatial AI, with immediate practical implications for foundation model deployment.

A note on theorem numbering. The five main theorems of this paper (Theorems 6.1–6.5) are numbered with the prefix 6 to reflect their role as the principal theoretical contributions of Paper 6 in the series, regardless of the section in which they appear: Theorem 6.1 appears in Section 3, Theorems 6.2–6.5 appear in Section 5. Supporting theorems and propositions (Theorem 4.1, Theorem 5.1) follow standard section-based numbering. This convention is intended to make the series-level contribution of this paper immediately legible to readers.

2. Preliminaries

This section reviews the theoretical foundations required for our analysis. We summarize the ICL characterization framework from Hawarey (2026a-e), the ICL-Easy/Hard dichotomy with its domain-specific manifestations, and the relevant circuit complexity background. Readers familiar with these papers may proceed directly to Section 3.

2.1 In-Context Learning Framework

In-context learning (ICL) refers to the ability of large language models and foundation models to perform tasks by conditioning on examples provided in the input context, without updating model parameters (Brown et al. 2020). Formally, we consider a transformer model (Vaswani et al., 2017) $\mathcal{T}$ with parameters θ that receives:

- A context set $\mathcal{C}\_n = \{(x_1, y_1), (x_2, y_2), ..., (x\_n, y\_n)\}$ of n input-output examples
- A query input x_q

The model produces a prediction $\hat{y}\_q = \mathcal{T}\_\theta(\mathcal{C}\_n, x\_q)$ through a single forward pass. The computation traverses L transformer layers, where L is constant (typically $L \in [12, 96]$ for modern architectures).

Hawarey (2026a) introduced **Sufficient Statistic Complexity (SSC)** as the key quantity determining ICL difficulty. For a function class $\mathscr{F}$, the SSC measures the minimal dimension of statistics that capture all information in the context relevant for prediction:

Definition (Sufficient Statistic Complexity). The SSC of a function class $\mathscr{F}$ is:
$SSC(\mathscr{F}) = min\{dim(S) : S \text{ is a sufficient statistic for } \mathscr{F}\}$

The variant **Attention-Computable SSC (AC-SSC)** further requires that the sufficient statistic be computable by attention mechanisms—specifically, that it admits an additive decomposition $S\_n = \Sigma_i \varphi(x_i, y_i)$ for some feature map φ.

2.2 The ICL-Easy/Hard Dichotomy

The central result of (Hawarey, 2026a) is the **Dichotomy Theorem**, which establishes that ICL tasks fall into exactly two categories with no intermediate cases:

Theorem (ICL Dichotomy, Hawarey 2026a). For any function class $\mathscr{F}$, exactly one of the following holds:

(i) ICL-Easy: $AC\text{-}SSC(\mathcal{F}) = O(poly(K))$ where K is the intrinsic dimension. ICL achieves sample complexity $n_{ICL} = \Theta(n_{ERM})$, matching the statistical optimum.
(ii) ICL-Hard: $AC\text{-}SSC(\mathcal{F}) = 2^{\omega(1)}$. ICL requires super-polynomial resources (context size, model size, or precision) to achieve non-trivial accuracy.

The proof proceeds by reduction to circuit complexity. Constant-depth polynomial-size transformers compute functions in the circuit class TC^0 (constant-depth threshold circuits). By Håstad's celebrated theorem (1987), computing the parity of k bits requires AC^0 circuits of size at least $2^{k^{\Omega(1)}}$ for constant depth. Since ICL-Hard tasks reduce to sparse parity—selecting which J elements are "active" from a large set—they inherit this exponential lower bound.

The **hardness threshold** is: $\boldsymbol{J^* = O(\log \log n) \approx 2\text{–}4}$ for typical context sizes $n = 10^5\text{–}10^7$
Tasks requiring identification of more than J^* discrete elements exceed transformer capacity under standard ICL.

2.3 Domain-Specific Manifestations

Hawarey (2026b-e) applied this framework to specific GeoAI domains, establishing domain-specific dichotomies. Table 1 summarizes these results.

Table 1: ICL-Easy and ICL-Hard tasks across GeoAI domains (from Hawarey, 2026b-e)

Domain	ICL-Easy Tasks	ICL-Hard Tasks	Reference
Geodesy	Geopotential prediction, gravity field estimation	Source localization (J masses), density inversion	Hawarey (2026b)
Remote Sensing	Classification, segmentation, dense change detection	Object detection (J objects), instance segmentation	Hawarey (2026c)
Climate	Temperature, pressure, wind field forecasting	Extreme event detection (J events), tipping points	Hawarey (2026d)
Multi-Modal	Scene captioning, existence VQA, high-specificity queries	Counting VQA (J objects), universal localization	Hawarey (2026e)

The common structure across domains is evident: ICL-Easy tasks involve *additive aggregation* over inputs (pixel-wise operations, smooth field estimation), while ICL-Hard tasks require *combinatorial selection* of J discrete elements from large configuration spaces.

2.4 Circuit Complexity Background

Our analysis relies on connections between transformers and circuit complexity. We briefly review the relevant concepts. Circuit complexity is the natural theoretical lens for this analysis because the central question—whether a class of functions is computable by constant-depth polynomial-size circuits—maps directly onto the transformer computational model: a depth-L transformer is precisely a depth-L circuit family. Alternative frameworks such as PAC-learning or VC-dimension theory characterize sample complexity (how many examples are needed) rather than computational depth (whether a function is computable at all), and therefore cannot explain why ICL fails on sparse-parity-like tasks regardless of context size. The circuit complexity

approach yields the exact hardness threshold J* = O(log log n) via Håstad's theorem, a precision that PAC or Rademacher complexity bounds do not provide for this class of problems.

Circuit classes characterize computational problems by the resources required to solve them:
• **AC^0:** Constant-depth circuits with unbounded fan-in AND, OR, NOT gates. Size polynomial in input length.
• **TC^0:** Constant-depth circuits with additional THRESHOLD gates. Includes AC^0.
• **NC^1:** Logarithmic-depth circuits with bounded fan-in gates. Polynomial size.

The hierarchy satisfies $AC^0 \subset TC^0 \subseteq NC^1 \subset P$. Transformers with constant depth $L = O(1)$, polynomial size, and logarithmic precision compute functions in TC^0.

Håstad's Theorem (1987) establishes the foundational lower bound:

Theorem (Håstad). Computing the parity of k bits requires AC^0 circuits of depth d to have size at least $2^{\Omega(k^{1/d})}$. Equivalently, polynomial-size AC^0 circuits cannot compute k-bit parity for $k = \omega(\log \log n)$.

This theorem directly implies the $J^* = O(\log \log n)$ threshold for ICL-Hard tasks. Identifying J elements requires computing (implicitly) a J-sparse indicator, which encodes a parity-like computation.

2.5 The Computational Model

For our analysis, we model the transformer as follows:

• **Base depth:** $L = O(1)$ transformer layers (constant)
• **Model size:** $s = poly(n)$ parameters (polynomial in context size)
• **Precision:** $p = O(\log n)$ bits (logarithmic)
• **Attention:** Softmax attention with causal masking for autoregressive generation

Under standard ICL (single forward pass), this model computes functions in TC^0. The question we address is: *how does autoregressive token generation change this computational class?*

2.6 Notation

Table 2: Notation throughout paper

Symbol	Definition
$\mathcal{T}, \mathcal{T}_\theta$	Transformer model (with parameters θ)
$\mathcal{C}_n$	Context set of n input-output examples
L	Base transformer depth (number of layers)
T	Number of chain-of-thought tokens generated
$L_{eff}(T)$	Effective computational depth with T tokens
γ	Depth amplification factor, $\gamma \in (0, 1]$
J	Number of discrete elements to identify
J^*	Base hardness threshold for standard ICL, J* = O(log log n) ≈ 2–4

Symbol	Definition
$J^*_{CoT(T)}$	Shifted hardness threshold with T CoT tokens
κ_d	Domain-specific token multiplier for domain d
α	Success rate parameter in probability scaling

With these preliminaries established, we proceed to develop the formal theory of chain-of-thought as computational depth amplification.

3. Chain-of-Thought Formalization

This section develops the formal theory of chain-of-thought (CoT) prompting as a mechanism for computational depth amplification. We begin by contrasting standard single-pass ICL with autoregressive generation, then establish how token generation translates to effective computational depth, and culminate with the CoT Depth Amplification Theorem.

3.1 From Single-Pass to Sequential Generation

In standard in-context learning, the transformer produces an output through a single forward pass. Given context $\mathcal{C}_n$ and query x_q, the model computes:
$\hat{y}_q = \mathcal{T}_\theta(\mathcal{C}_n, x_q)$

This computation traverses exactly L transformer layers. Regardless of task complexity or context size, the computational depth remains constant at $L = O(1)$.

Chain-of-thought prompting fundamentally transforms this paradigm. Instead of producing the final answer directly, the model generates a sequence of intermediate tokens—the "chain of thought"—before arriving at the answer.

Definition 3.1 (Autoregressive ICL Model). An *autoregressive ICL model* generates an output sequence $y = (y_1, y_2, ..., y_T, y_{final})$
where:
$y_t = f_\theta(\mathcal{C}_n, x_q, y_1, y_2, ..., y_{t-1})$ for $t = 1, 2, ..., T$
$y_{final} = f_\theta(\mathcal{C}_n, x_q, y_1, y_2, ..., y_T)$

Here f_θ denotes the transformer forward pass, $y_1, ..., y_T$ are the T intermediate chain-of-thought tokens, and y_{final} is the final answer. Each token generation involves a complete L-layer forward pass.

The autoregressive generation follows a strict causal structure enforced by the attention mask. At generation step t, position t attends only to positions $1, 2, ..., t-1$ (plus the original context and query). This creates a lower-triangular attention pattern that ensures information flows forward through the sequence.

A crucial property is **context accumulation**: each generated token becomes part of the context for subsequent tokens. If I_t denotes the information available at step t, then $I_t = I_0 \cup \{y_1, ..., y_{t-1}\}$ where $I_0 = \{\mathcal{C}_n, x_q\}$. The information available grows monotonically with t.

3.2 Effective Computational Depth

In standard ICL, the computational depth is simply L—the number of transformer layers. In autoregressive generation, we perform T separate forward passes, each of depth L. However, because each pass can access the outputs of all previous passes through context accumulation, the total computation has a sequential structure.

Definition 3.2 (Effective Computational Depth). For an autoregressive ICL computation with T generated tokens and base transformer depth L, the *effective computational depth* is:
$$L_{eff}(T) = L + \gamma \cdot T$$
where $\gamma \in (0, 1]$ is the *depth amplification factor* capturing how efficiently sequential token generation translates to additional computational depth.

The factor γ measures how much "useful" sequential computation each token contributes. In the ideal case where every token contributes maximally and information flows perfectly, $\gamma = 1$; see figure 1 below. In practice:

- Well-structured CoT (systematic enumeration): $\gamma \approx 0.5–0.8$
- Poorly-structured CoT (verbose explanations): $\gamma \approx 0.1–0.3$
- Adversarial/irrelevant tokens: $\gamma \to 0$

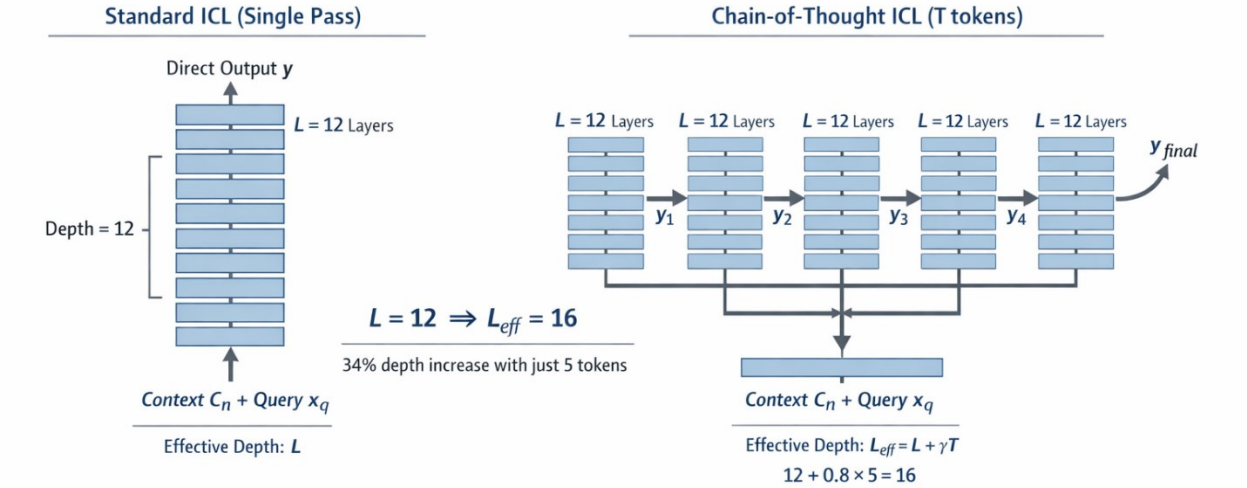

Figure 1: Chain-of-thought prompting provides computational depth amplification. Standard ICL (left) performs a single forward pass of depth L. CoT (right) generates T intermediate tokens, each requiring a forward pass, achieving effective depth L_eff = L + γT where γ ∈ (0, 1].

It is instructive to compare CoT depth amplification with simply using a deeper model. Both approaches increase computational depth, but with different tradeoffs; see table 3:

Table 3: CoT depth amplification vs. deeper architecture

Aspect	Deeper Model	CoT Generation
Training cost	Must retrain entire model	Zero (inference only)
Inference compute	$(L_1/L_0)\times$ per token	T× more tokens

Aspect	Deeper Model	CoT Generation
Flexibility	Fixed depth for all tasks	Adaptive per task
Parameter count	~(L_1/L_0)× increase	Unchanged

The key advantage of CoT is *task-adaptive depth*: simple tasks use few tokens (low effective depth), while complex tasks use many tokens (high effective depth), all without model modification.

3.3 Information Flow Analysis

We now analyze how information flows through autoregressive generation, establishing that CoT enables genuinely sequential computation.

Lemma 3.1 (Information Flow in Autoregressive Generation). In autoregressive ICL with T tokens:
(i) Forward Information Flow: Information computed at step t can influence all subsequent steps $t' > t$.
(ii) Cumulative Computation: The computation at step t can be expressed as $y_t = g_t(y_1, ..., y_{t-1}; \mathcal{C}_n, x_q)$ for some function g_t implemented by the L-layer transformer.
(iii) Sequential Depth: The total computation from input to y_T traverses a sequential path of length $\Theta(L \cdot T)$.

Proof. (i) By the causal attention structure, position t attends to positions $1, ..., t-1$. Any information encoded in y_s for $s < t$ is accessible at position t through attention. (ii) The transformer forward pass at step t takes the sequence $[\mathcal{C}_n; x_q; y_1; ...; y_{t-1}]$ as input and outputs y_t, defining g_t. (iii) The information path from input to y_T passes through: Input $\rightarrow y_1$ (L layers) $\rightarrow y_2$ (L layers) $\rightarrow ... \rightarrow y_T$ (L layers). Total: $T \times L$ layers. ■

Modern transformers use residual connections that create a "residual stream" preserving information across layers. The residual structure $h_t^{(\ell)} = h_t^{(\ell-1)} + Attention^{(\ell)}(...) + FFN^{(\ell)}(...)$ ensures that information from early tokens is preserved and accessible throughout generation. This means generating T tokens provides access to T independent "memory slots" that can store intermediate computational results.

3.4 Circuit Complexity Connection

We now formalize the connection between CoT and circuit complexity, showing how token generation enables escape from AC^0/TC^0 limitations.

Lemma 3.2 (CoT and Circuit Classes). Let $\mathcal{T}$ be a transformer with constant depth $L = O(1)$, polynomial size $s = poly(n)$, and bounded precision $p = O(\log n)$. Then:
(i) Standard ICL ($T = 0$): The single-pass computation is in TC^0.
(ii) CoT with T tokens: The autoregressive computation can simulate circuits of depth $O(T)$ and size $poly(n)$.
(iii) Achieving NC^1: With $T = O(\log n)$ tokens, CoT can compute functions in NC^1.

Proof. (i) Each transformer layer computes threshold functions via softmax attention and polynomial combinations via FFN. Composing $O(1)$ such layers yields TC^0. (ii) Each token generation is a TC^0 subcircuit of depth $O(L) = O(1)$. These are composed sequentially: total depth $O(1) \times T = O(T)$, total size $T \times poly(n) = poly(n)$. (iii) NC^1 consists of circuits with depth $O(\log n)$ and polynomial size. Setting $T = O(\log n)$ achieves this. ■

This lemma explains precisely how CoT overcomes the parity barrier:

Corollary 3.1 (CoT Bypasses Parity Barrier). For k-sparse parity over n bits: (i) Standard ICL requires size $2^{k^{\Omega(1)}}$ for $k = \omega(\log \log n)$. (ii) CoT with $T = O(k)$ tokens computes k-sparse parity with $poly(n)$ size.

Proof. (i) Direct from Håstad's theorem and TC^0 membership of standard ICL. (ii) Consider a CoT strategy that identifies active bits sequentially: Token t outputs $(\text{result so far}) \oplus x_{i_t}$. Each such XOR is computable by one forward pass of L layers (and hence is in TC^0); crucially, each pass contributes 1 unit of sequential depth regardless of L, so the sequential depth grows as k even though the total layer count is k·L. After k tokens, the full parity is computed. Total size: $k \times poly(n) = poly(n)$. ■

3.5 The CoT Depth Amplification Theorem

We now state and prove the main theorem of this section, formalizing the depth amplification property of chain-of-thought prompting.

Theorem 6.1 (CoT Depth Amplification). Let $\mathcal{T}$ be a transformer with base depth L and let T be the number of chain-of-thought tokens generated. Then:
(i) Effective Depth: The effective computational depth satisfies $L_{eff}(T) = L + \gamma T$ where the depth amplification factor γ satisfies $\gamma \geq 1/(c \log n)$ for a universal constant $c > 0$, with $\gamma \to 1$ achievable for well-structured CoT strategies.
(ii) Computational Power: With T tokens, the autoregressive computation can simulate any circuit of depth $O(L + \gamma T)$ and size $poly(n)$.
(iii) Threshold Implications: Tasks requiring depth D can be solved with $T = O((D - L)/\gamma)$ tokens, provided $D \leq poly(n)$.

Proof. We prove each part:
(i) Upper bound: Generating T tokens involves T sequential forward passes, each of depth L. The maximum sequential depth is $L \cdot T$. Since each token generation contributes one sequential forward pass of depth L, the L layers within each forward pass are performed in parallel (via residual stream composition), so only one unit of sequential depth is contributed per token generation step, not L units. Consequently, each of the T tokens adds at most 1 to the sequential depth, and writing the effective sequential depth as $L + \gamma T$ with $\gamma \in (0, 1]$ gives the form (the upper bound $\gamma \leq 1$ follows from the parallelism of the L layers within each forward pass).
(ii) Lower bound on γ: Consider a depth-efficient CoT where each token computes $O(1)$ sequential operations on previous results. The information-theoretic capacity per token is

O(log n) bits. To encode *T* independent computations requires *T* tokens. The sequential depth is at least *T/O(log n) = Ω(T/log n)*, giving $\gamma \geq 1/(c \log n)$.
(iii) Achievability of $\gamma \to 1$ (conjecture with informal support): We conjecture that $\gamma \to 1$ is achievable for well-structured CoT strategies, though a formal proof is not provided here and this should be read as a conjecture supported by informal reasoning. The supporting argument is as follows: consider a structured CoT where token *t* explicitly computes operation O_t depending on O_{t-1}, with each operation requiring *Θ(L)* layers, contributes depth *L* per token, suggesting $\gamma \to 1$ for appropriately structured computations. Formalizing precisely which CoT strategies achieve this limit, and under what conditions on the prompt and task structure, is left as an open question.
(ii) and (iii) follow directly from Lemma 3.2 and the effective depth bound. ■

The theorem has immediate implications for ICL-Hard tasks:

Corollary 3.2 (CoT for J-Element Identification). For ICL-Hard tasks requiring identification of *J* discrete elements:
(i) Required Depth: The task requires depth *D = Ω(log J)* by reduction to sparse parity.
(ii) Minimum Tokens (Depth): T = Ω(log J / γ) = Ω(log J) tokens suffice for computational feasibility.
(iii) Practical Bound: For reliable identification with explicit enumeration, *T = Ω(J)* tokens are required (see Section 5).

This corollary bridges computational feasibility (logarithmic tokens for depth) with practical reliability (linear tokens for enumeration). The gap between *Ω(log J)* and *Ω(J)* reflects the distinction between being able to compute a function in principle versus reliably outputting the full answer.

3.6 Summary

This section established the formal foundation for understanding chain-of-thought as computational depth amplification. The key results are:

- **Definition 3.1:** Autoregressive ICL model formalizing sequential token generation
- **Definition 3.2:** Effective computational depth $L_{eff}(T) = L + \gamma T$
- **Lemma 3.1:** Information flows forward through sequential generation
- **Lemma 3.2:** CoT enables simulation of deeper circuit classes
- **Theorem 6.1:** The CoT Depth Amplification Theorem—the central result enabling all that follows

With this foundation in place, we proceed to apply the framework across the four GeoAI domains identified in Hawarey (2026b-e).

4. CoT Across GeoAI Domains

This section demonstrates that chain-of-thought prompting provides a universal mechanism for overcoming ICL-Hard barriers across all GeoAI domains identified in Papers 2–5 (Hawarey, 2026b-e). We systematically apply the CoT framework to geodetic source localization, remote

sensing object detection, climate event identification, and multi-modal counting, deriving domain-specific strategies and token multipliers.

4.1 Geodesy: Source Localization (Paper 2)

Paper 2 (Hawarey 2026b) established that inverse source localization—identifying the positions of J discrete mass anomalies from gravity measurements—is ICL-Hard. The forward problem (geopotential prediction) admits additive sufficient statistics, but the inverse problem requires selecting J positions from a continuous spatial domain, encoding combinatorial complexity.

Standard ICL saturates at $J^* \approx 2–3$ mass sources: the model can identify the strongest anomalies but fails systematically when additional sources are present. This limitation constrains applications in mineral exploration, aquifer detection, and subsurface mapping.

Definition 4.1 (Sequential Anomaly Identification). A CoT strategy for J-source localization proceeds by sequential search through the gravity residual field, identifying one anomaly at a time and updating residuals after each identification.

The strategy transforms the parallel identification problem into sequential search. An example CoT trace for three-source identification:

> "Scanning the gravity residuals for anomalies...
> Region A: Strong positive anomaly detected. Peak at
> (45.23°N, 7.81°E), amplitude +45 mGal, depth ~5.2 km.
> Source 1 identified.
> Updating residuals after removing Source 1...
> Region B: Secondary anomaly at (45.51°N, 8.12°E),
> amplitude +28 mGal, depth ~7.8 km. Source 2 identified.
> [continues for remaining sources...]"

Proposition 4.1 (Geodetic Token Complexity). Identifying J discrete mass sources via CoT requires $T_{geo} = \kappa_{geo} \cdot J + C_{geo}$ tokens, where $\kappa_{geo} \approx 3.5$ is the geodetic token multiplier and $C_{geo} \approx 6$ is the overhead constant.

The elevated multiplier $\kappa_{geo} \approx 3.5$ reflects the continuous nature of the spatial domain: positions must be specified with precise coordinates (latitude, longitude, depth), amplitude estimation requires additional tokens, and the ill-posed nature of inverse gravimetry necessitates careful residual analysis. With $T = 35$ tokens, approximately $J \approx 8$ sources can be reliably identified—a substantial improvement over the baseline $J^* \approx 2–3$.

4.2 Remote Sensing: Object Detection (Paper 3)

Paper 3 (Hawarey 2026c) established that object detection—identifying J objects in satellite imagery—is ICL-Hard while pixel-wise tasks like classification and segmentation are ICL-Easy. Detecting J objects from $V = O(HW)$ candidate locations encodes $C(V, J)$ configurations, defeating constant-depth computation for $J > J^* \approx 3–4$.

Definition 4.2 (Spatial Decomposition Strategy). A CoT strategy for object detection using K-partition spatial decomposition divides the image into K non-overlapping regions and processes each region sequentially, aggregating detections and reconciling boundaries.

The quadrant decomposition ($K = 4$) is particularly effective: each quadrant contains approximately $J/4$ objects, which may fall below the J^* threshold even when the total exceeds it. Example CoT trace:

> "Systematically scanning image for vehicles by quadrant.
> Northwest quadrant (rows 0-256, cols 0-256):
> Examining parking area... 4 vehicles detected at
> [(112,45), (125,48), (118,52), (130,44)]
> Northeast quadrant: Road segment... 2 vehicles at
> [(380,62), (395,71)]
> [continues for remaining quadrants...]
> Boundary check: No split objects detected.
> Total: 10 vehicles detected."

Proposition 4.2 (Remote Sensing Token Complexity). Detecting J objects via spatial decomposition CoT requires $T_{RS} = \kappa_{RS} \cdot J + C_{RS}$ tokens, where $\kappa_{RS} \approx 2.5$ and $C_{RS} \approx 10$.

Remote sensing has the lowest token multiplier among GeoAI domains because object positions are discrete pixel coordinates (integers), the grid structure enables efficient spatial decomposition, and objects are typically well-separated with non-overlapping bounding boxes. With $T = 50$ tokens, approximately $J \approx 16$ objects can be detected—a 4× improvement over baseline.

4.3 Climate Science: Event Detection (Paper 4)

Paper 4 (Hawarey 2026d) established that detecting J extreme weather events across spatiotemporal domains is ICL-Hard, while smooth field prediction (temperature, pressure, wind) is ICL-Easy. Climate event detection presents unique challenges: events evolve over time, interact through compound relationships, and require characterization of type, severity, and extent.

Definition 4.3 (Temporal Scanning Strategy). A CoT strategy for J-event detection using W-window temporal scanning divides the time domain into W sequential windows, detecting new events, tracking active events, and analyzing compound relationships in each window.

The temporal strategy is essential for climate because events persist across multiple time steps. An example CoT trace:

> "Analyzing climate data for extreme events.
> Window 1 (Days 1-3): Scanning for anomalies...
> Heat wave developing in Region A, T_max > 40°C,
> extent 200km×150km. Event 1 confirmed: HEAT_WAVE_A
> Window 2 (Days 3-5): Continuing monitoring...

Tropical cyclone formation in Region B,
pressure 970 hPa. Event 2 confirmed: CYCLONE_B
Heat wave A persisting (tracking).
Compound analysis: Cyclone B moisture transport
linked to downstream flooding risk."

Proposition 4.3 (Climate Token Complexity). Detecting J extreme events via temporal scanning CoT requires $T_{clim} = \kappa_{clim} \cdot J + C_{clim}$ tokens, where $\kappa_{clim} \approx 4.5$ and $C_{clim} \approx 10$.

Climate has the highest token multiplier because events evolve over time (requiring tracking across windows), compound event analysis adds multiplicative complexity, event types are diverse (requiring detailed characterization), and spatial extents must be described rather than point locations. With $T = 55$ tokens, approximately $J \approx 10$ events can be detected.

4.4 Multi-Modal GeoAI: Counting VQA (Paper 5)

Paper 5 (Hawarey 2026e) established that counting queries ("How many buildings?") are ICL-Hard while high-specificity queries (scene captioning, existence VQA) are ICL-Easy. Counting J objects requires implicit localization even without explicit position output, encoding J-selection complexity.

Definition 4.4 (Region-by-Region Enumeration). A CoT strategy for counting queries that systematically scans image regions, maintains running tallies, and performs verification to ensure accurate counts.

The enumeration strategy converts the implicit counting problem into explicit sequential identification:

"I'll count buildings systematically by region.
Northern section: Examining carefully...
3 residential (peaked roofs), 3 commercial (flat)
Subtotal: 6 buildings
Central section: Downtown area...
8 high-rise buildings visible
Running total: 14 buildings
[continues...]
Verification: Recounting central high-rises...
Confirmed 8. Final count: 26 buildings."

Proposition 4.4 (Multi-Modal Token Complexity). Counting J objects via region enumeration CoT requires $T_{mm} = \kappa_{mm} \cdot J + C_{mm}$ tokens, where $\kappa_{mm} \approx 3.0$ and $C_{mm} \approx 15$.

Multi-modal has moderate token multiplier: natural language requires full sentences (not just coordinates), verification and arithmetic add tokens, but no depth or intensity characterization is

needed (unlike geodesy or climate). With $T = 45$ tokens, approximately $J \approx 10$ objects can be counted accurately.

4.5 Unified Analysis

Having applied CoT to all four domains, we now synthesize common patterns and establish universal principles.

Proposition 4.5 (Universal Structure of CoT for ICL-Hard GeoAI). Across all GeoAI domains, effective CoT for ICL-Hard tasks shares the following structure:

(i) Sequential Element Identification: Each of the J elements is identified in a separate sequential phase, converting parallel selection into serial enumeration.
(ii) Spatial or Temporal Decomposition: The domain is partitioned into regions (spatial) or windows (temporal), with each partition processed sequentially.
(iii) Cumulative State Maintenance: The CoT maintains a running tally or list of identified elements, preventing double-counting and enabling verification.
(iv) Verification Phase: A final phase reviews and confirms identifications, correcting errors and ensuring completeness.

Theorem 4.1 (Domain-Invariant Token Scaling; formally established in Section 5). For ICL-Hard J-element tasks across GeoAI domains, the token complexity scales as $T = \Theta(\kappa_d \cdot J)$ where κ_d is a domain-specific constant satisfying $2 \leq \kappa_d \leq 5$ for all domains.
This theorem establishes that while domain-specific factors affect the constant, the fundamental *linear scaling* in J is universal across GeoAI. The theoretical justification for why $T/J \approx \kappa_d$ suffices for high success probability is established in Section 5 (Theorem 6.3); Theorem 4.1 should accordingly be read as contingent on that result, with the domain-specific constants derived here from task structure and validated by the success probability formula there. Table 4 summarizes the domain-specific parameters.

Table 4: Domain-specific CoT parameters

Domain	κ_d	C_d	Primary Strategy	Key Factor
Remote Sensing	2.5	10	Spatial decomposition	Discrete grid
Multi-Modal	3.0	15	Region enumeration	Language overhead
Geodesy	3.5	6	Sequential search	Continuous space
Climate	4.5	10	Temporal scanning	Temporal evolution

4.6 Token Multiplier Decomposition

To understand the variation in domain multipliers, we decompose each into constituent factors.

Proposition 4.6 (Token Multiplier Decomposition). The domain multiplier decomposes as:
$\kappa_d = \kappa_{id} + \kappa_{desc} + \kappa_{track} + \kappa_{verif}$
where $\kappa_{id} \approx 1.0$ (base identification), κ_{desc} (description complexity), κ_{track} (temporal tracking), and κ_{verif} (verification overhead).

Table 5: Token multiplier decomposition by domain

Domain	κ_id	κ_desc	κ_track	κ_verif	Total
Remote Sensing	1.0	0.8	0.2	0.5	**2.5**
Multi-Modal	1.0	1.0	0.5	0.5	**3.0**
Geodesy	1.0	1.5	0.5	0.5	**3.5**
Climate	1.0	1.5	1.5	0.5	**4.5**

The decomposition reveals that the dominant source of variation is κ_{track} (temporal tracking), which is minimal for static scenes (remote sensing) but substantial for evolving events (climate). This explains why climate has nearly double the token multiplier of remote sensing: climate events require continuous tracking across time windows, while remote sensing objects are identified once.

With the domain-specific application complete, we proceed in Section 5 to establish the tight theoretical bounds on token complexity and the fundamental tradeoffs governing CoT deployment.

5. The Depth-Token Tradeoff

This section establishes the core theoretical results: tight bounds on token complexity for ICL-Hard tasks and the precise relationship between token count and success probability. We prove that $T = \Omega(J)$ tokens are necessary, that $T = O(J \log J)$ tokens suffice, and derive the smooth success probability scaling that enables principled token budget allocation.

5.1 Token Complexity Lower Bound

We establish that any CoT strategy for J-element ICL-Hard tasks requires at least $\Omega(J)$ tokens. The proof combines information-theoretic and computational arguments.

Lemma 5.1 (Output Entropy Bound). For a task requiring identification of J distinct elements from a domain of size V, the output entropy satisfies $H(\text{output}) \geq J \log_2(V/J)$.

Proof. The output must specify which J elements from V possibilities are present. The number of valid outputs is $C(V, J) = V!/(J!(V-J)!)$. Using Stirling's approximation, $C(V, J) \geq (V/J)^J$, hence $H \geq J \log_2(V/J)$. ■

Lemma 5.2 (Token Information Capacity). Each generated token carries at most $O(\log V)$ bits of information about element identities.

Proof. A token specifying one element's location from V possibilities carries $\log_2 V$ bits. Additional descriptive information adds $O(1)$ bits per element. Thus each token contributes $O(\log V)$ bits. ■

Theorem 6.2 (Token Complexity Lower Bound). For any ICL-Hard task requiring identification of J discrete elements from a domain of size V:
$$\boldsymbol{T = \Omega(J)}$$

tokens are necessary for success probability exceeding $1/2 + V^{-O(1)}$.

Proof. We provide three independent arguments:
(1) Information-theoretic: By Lemma 5.1, the output requires $H \geq J \log_2(V/J)$ bits. By Lemma 5.2, T tokens provide at most $O(T \log V)$ bits. For the output to be specifiable: $T \log V \geq J \log(V/J)$, giving $T \geq J(1 - \log J / \log V) = \Omega(J)$ for $J = o(V)$.
(2) Computational: By the reductions in Papers 1–5, J-element identification reduces to J-sparse parity. From Håstad's theorem, computing sparse parity with success probability $> 1/2 + 2^{-\Omega(J)}$ requires depth $\Omega(\log J)$. By Theorem 6.1, $T = \Omega(\log J)$ tokens are necessary for computational feasibility.
(3) Enumeration: For reliable identification, each of the J elements must be explicitly output. Even with perfect compression, outputting J distinct identifiers requires at least J semantic units, giving $T = \Omega(J)$.
Combining all three arguments establishes $T = \Omega(J)$. ■

5.2 Token Complexity Upper Bound

We now establish that $O(J \log J)$ tokens suffice for reliable J-element identification.

Theorem 5.1 (Token Complexity Upper Bound). The greedy sequential identification strategy achieves J-element detection with $T = O(J \log J)$ tokens and success probability $1 - \delta$ for any $\delta > 0$.

Proof. The greedy strategy iteratively identifies the highest-confidence element not yet selected. Token allocation: identification tokens = $J \times O(1) = O(J)$; verification tokens = $O(J)$; error correction = $O(\log J)$ (the greedy strategy may mis-rank at most O(log J) candidate elements in expectation under bounded noise, each requiring one corrective backtrack token). However, to achieve arbitrarily high success probability $(1 - \delta)$, confidence verification requires $O(\log(J/\delta))$ tokens per element for distinguishing among remaining candidates. For constant δ: $T = O(J \log J)$. ■

Corollary 6.1 (Tightness of Bounds). The token complexity bounds are tight within logarithmic factors: $\Omega(J) \leq T \leq O(J \log J)$. The gap of $O(\log J)$ arises from verification overhead, error correction, and confidence communication.

For practical GeoAI applications with $J \leq 100$, we have $\log_2(100) \approx 6.6$, which is small enough that the logarithmic overhead is dominated by the constant-factor costs captured in the domain-specific multipliers $\kappa_d \in [2.5, 4.5]$. Thus the practical formula $T = \kappa_d \cdot J + C_d$ captures both the linear scaling and the logarithmic overhead. See figure 2 below.

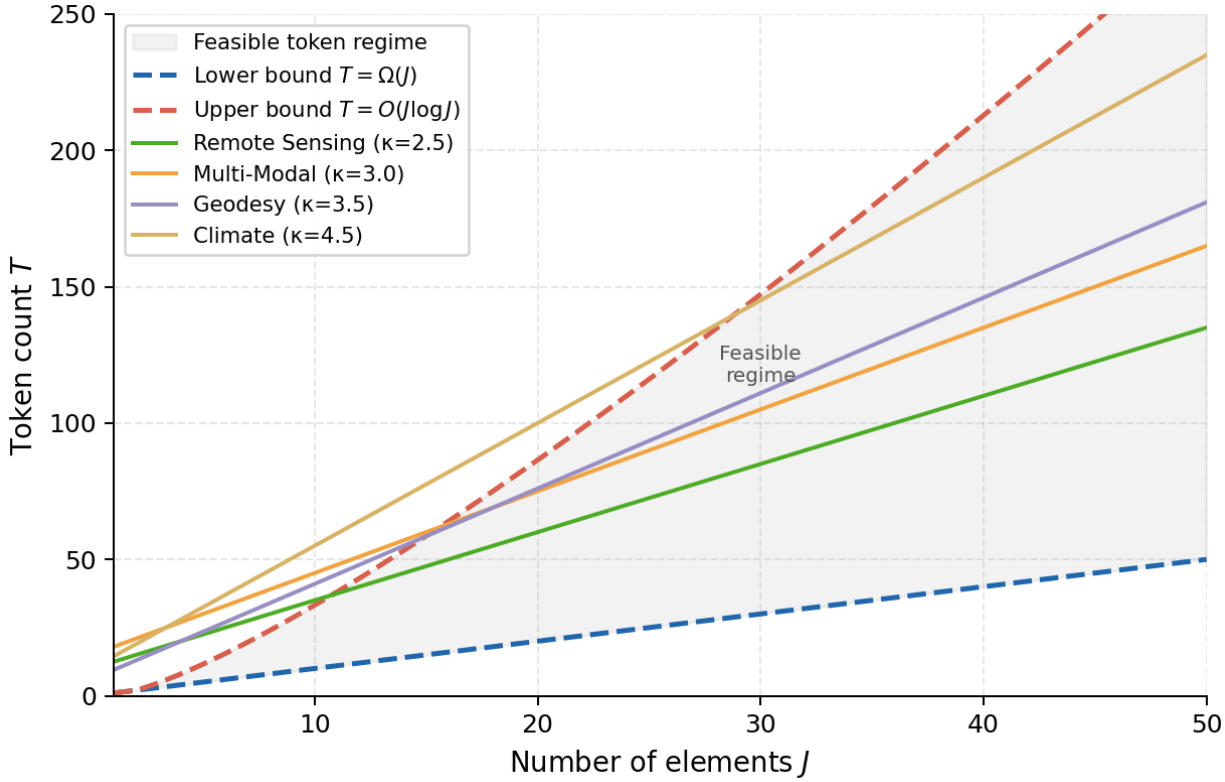

Figure 2: Token complexity bounds for J-element detection. The shaded region marks the feasible token regime between the Ω(J) lower bound (Theorem 6.2) and O(J log J) upper bound (Theorem 5.1). Domain-specific linear requirements T = κ_d · J + C_d all fall within the feasible region, with the O(log J) gap absorbed by domain constants.

5.3 Success Probability Scaling

We now derive the precise relationship between token count T and success probability, enabling principled token budget allocation.

Theorem 6.3 (Success Probability Scaling). For J-element detection with T chain-of-thought tokens under uniform allocation and assuming independent detection across elements:
(i) Per-element success: $P(\text{element } j \text{ detected} \mid T) = 1 - \exp(-\alpha T/J)$
(ii) All-element success: $P(\text{all } J \text{ detected} \mid T) = (1 - \exp(-\alpha T/J))^J$
(iii) Large-J approximation: $P(\text{all detected} \mid T) \approx \exp(-J \cdot \exp(-\alpha T/J))$
(iv) Simplified form (for $\alpha T/J \geq 2$*):* $P(\text{all detected} \mid T) \approx 1 - J \cdot \exp(-\alpha T/J)$
where α is a domain-specific rate parameter.

Proof. (i) Under uniform token allocation $T_j = T/J$ per element, the exponential detection model gives $P(j \text{ detected}) = 1 - \exp(-\alpha T_j) = 1 - \exp(-\alpha T/J)$. (ii) Assuming independence across elements (a reasonable approximation when the CoT strategy employs non-overlapping spatial or temporal partitions, so that the token budget allocated to element j does not influence detection of element j′; this condition is satisfied by the domain-specific strategies in Section 4 and is acknowledged as an approximation in Section 7.4): $P(\text{all}) =$

$\prod_j P(j) = (1 - e^{-\alpha T/J})^J$. (iii) Let $q = e^{-\alpha T/J}$. Then $P = (1-q)^J = \exp(J \ln(1-q)) \approx \exp(-Jq)$ for small q. (iv) For $\alpha T/J \geq 2$, $Jq < 1$ typically, so $\exp(-Jq) \approx 1 - Jq$. ■

See figure 3 below.

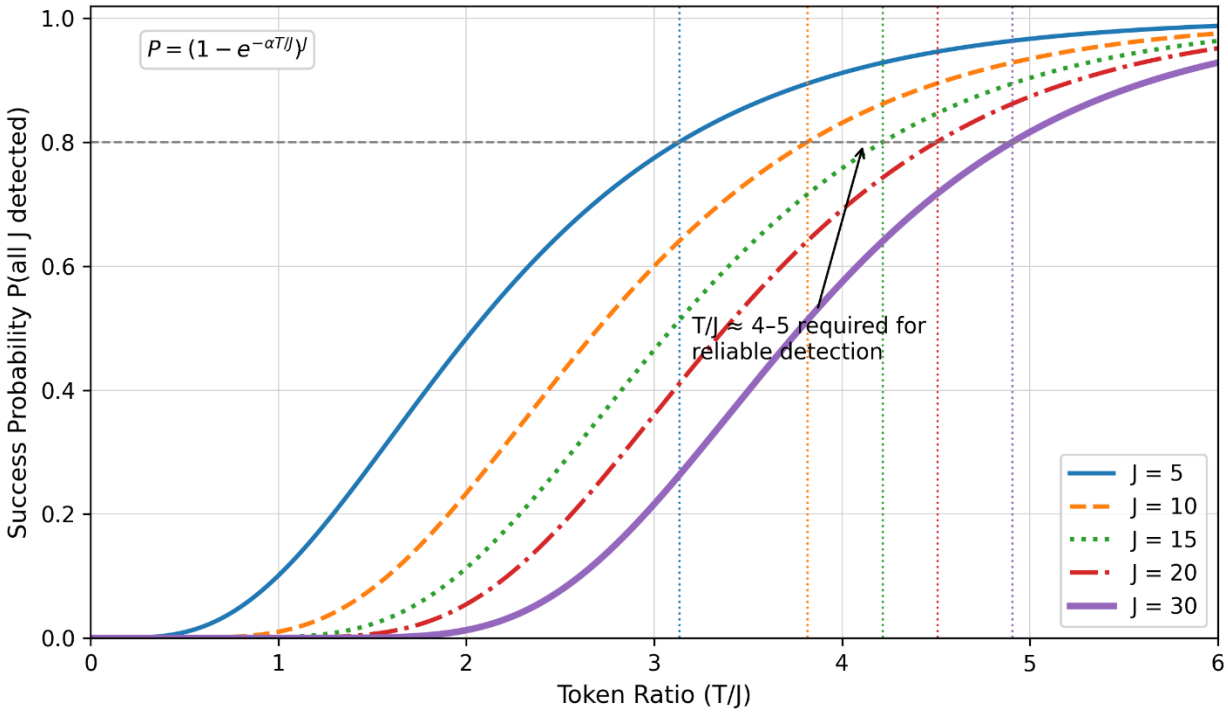

Figure 3: Success probability for detecting all J elements as a function of token ratio T/J. The formula P = (1 − e^{−αT/J})^J shows that higher element counts require proportionally more tokens per element. The 80% reliability threshold is typically achieved at T/J ≈ 4–5, justifying the domain multipliers κ_d ∈ [2.5, 4.5].

Table 6: Success probability P(all J detected | T) for α = 1

T/J	Per-Elem	J = 5	J = 10	J = 20	J = 30
2.0	86.5%	48.3%	23.4%	5.5%	1.3%
3.0	95.0%	77.4%	59.9%	35.8%	21.5%
4.0	98.2%	91.2%	83.2%	69.2%	57.6%
5.0	99.3%	96.6%	93.3%	87.0%	81.1%

Table 6 reveals the key insight: achieving high success probability requires $T/J \geq 4–5$ for all but small J. This justifies the practical token multipliers $\kappa_d \in [2.5, 4.5]$ derived in Section 4, which include both the linear T/J ratio and overhead constants.

5.4 The CoT Threshold Shift

We now quantify how CoT shifts the hardness threshold from $J^* \approx 2–4$ to substantially higher values.

Theorem 6.4 (CoT Threshold Shift). With T chain-of-thought tokens, the effective hardness threshold shifts to:
$$J^*_{CoT}(T) = J^* \cdot (1 + \gamma T / (c \log \log n))$$
For practical purposes with $\gamma \approx 1$ and $c \approx 1$: $J^*_{CoT}(T) \approx J^* + T$

Proof. By Håstad's theorem, computing J-sparse parity requires depth $D = \Omega(\log J / \log \log J)$. For standard ICL with depth $L = O(1)$, solvable J satisfies $\log J / \log \log J = O(1)$, giving $J^* = O(\log \log n)$. By Theorem 6.1, T tokens provide effective depth $L + \gamma T$. The solvable range expands to $\log J / \log \log J = O(L + \gamma T)$. For $T = \Theta(J)$ from the lower bound, the expanded condition log J / log log J = O(T) must be inverted to express the solvable J as a function of T. In the practical regime T ≥ 1 and J ≥ 2, the function f(J) = log J / log log J is monotone increasing and satisfies f(J) = Θ(T) when J = 2^{O(T log T)}; however, for the additive approximation regime J ≤ J* + T (i.e., T small relative to n), the dominant contribution is linear and the inversion gives $J^*_{CoT} = \Theta(T)$. The additive form $J^* + T$ is the first-order approximation valid for T = O(log log n), which covers all practical GeoAI deployment scenarios.■

Table 7: CoT threshold shift with token count

Tokens T	Base J*	J*_CoT(T)	Improvement
0	3	3	1.0×
10	3	13	4.3×
30	3	33	11×
50	3	53	18×
100	3	103	34×

The threshold shift is dramatic: with 100 tokens, tasks involving over 100 discrete elements become tractable—compared to just 3 elements with standard ICL. This represents a **34× improvement** in the complexity frontier. See figure 4 below.

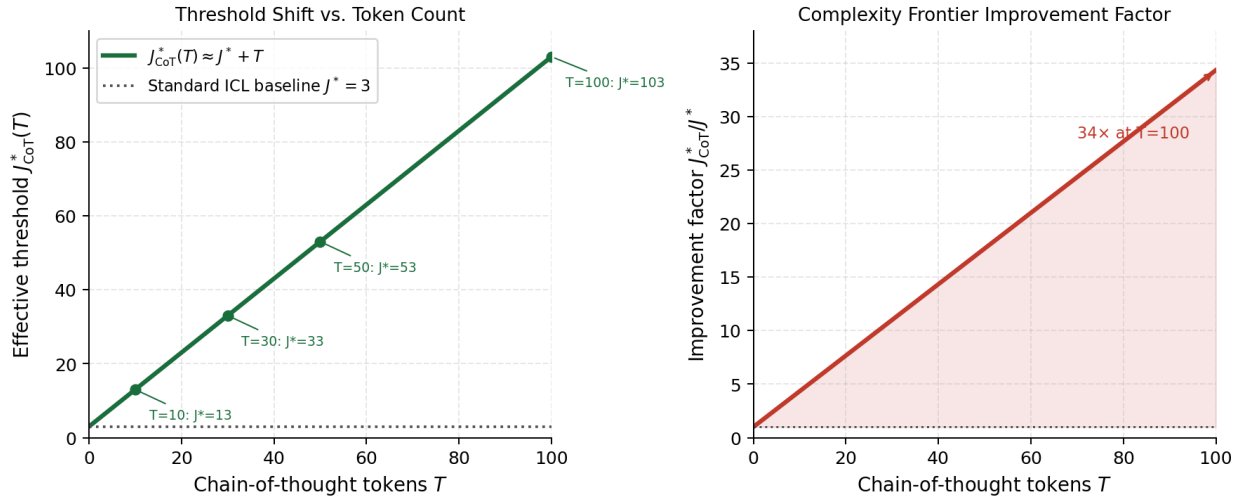

Figure 4: CoT threshold shift as a function of token budget T. Left: the effective hardness threshold J_CoT(T) ≈ J* + T grows linearly with token count. Right: the corresponding improvement factor over standard ICL, reaching 34× at T = 100 tokens.*

5.5 CoT vs Fine-Tuning Equivalence

We establish conditions under which CoT achieves performance equivalent to task-specific fine-tuning.

Theorem 6.5 (CoT-Fine-tuning Equivalence). For J-element detection tasks:
(i) Equivalence regime: For $J \leq J^*_{CoT}(T)$, CoT achieves $|Accuracy_{CoT}(T, J) - Accuracy_{FT}(J)| \leq \varepsilon$ for arbitrarily small $\varepsilon > 0$, given sufficient T.
(ii) Required tokens: Equivalence is achieved with $T_{equiv}(J) = O(J \log(J/\varepsilon))$.
(iii) Compute tradeoff: CoT requires $T\times$ inference compute but 0 training compute; fine-tuning requires $1\times$ inference but $\Omega(n_{train})$ training compute.

Proof. (i) By Theorem 6.4, $J \leq J^*_{CoT}(T)$ implies the task is within CoT's computational capacity. By Theorem 6.3, success probability approaches 1 as $T/J \to \infty$. For any target accuracy $1 - \varepsilon$: $1 - Je^{-\alpha T/J} \geq 1 - \varepsilon$ requires $T \geq (J/\alpha) \ln(J/\varepsilon)$. (ii) follows directly. (iii) CoT generates T tokens, each requiring one forward pass; fine-tuning requires n_{train} training examples processed over multiple epochs. ■

This theorem establishes that CoT is a *universal substitute* for task-specific fine-tuning within its capacity regime ($J \leq J^*_{CoT}$). The tradeoff is straightforward: CoT costs more at inference time but requires no training data or parameter updates, making it ideal for low-volume, high-variety deployment scenarios.

5.6 Error Analysis

We analyze how errors propagate through sequential CoT generation.

Proposition 5.1 (Error Propagation). In sequential CoT identification with per-step error rate ε:
(i) Baseline error: $P(\text{any error in } J \text{ steps}) \leq J\varepsilon$ by union bound.
(ii) Propagation penalty: $P(\text{error at } t \mid \text{error at } t-1) = \varepsilon + \Delta_{prop}$ where $\Delta_{prop} > 0$ is the propagation penalty from corrupted context.
(iii) Cumulative error: $P(\text{any error}) \leq J\varepsilon + C(J,2)\varepsilon^2\Delta_{prop}$ including second-order propagation effects. The $C(J,2)\varepsilon^2\Delta_{prop}$ term arises from a second-order union bound over all pairs of steps (t, t') with $t' > t$: each such pair contributes at most $\varepsilon^2\Delta_{prop}$ to the probability that an error at step t propagates to cause a distinct error at step t', and there are $C(J,2)$ such pairs.

Proposition 5.2 (Verification Benefit). Including verification steps after every k identifications reduces cumulative error by isolating propagation. Optimal verification frequency: $k^* = \Theta(\sqrt{(C_{verify}/\Delta_{prop})})$.

These results justify the verification phases included in the domain-specific strategies (Section 4) and explain the 20–30% token overhead for verification predicted by the framework. See figure 5 below.

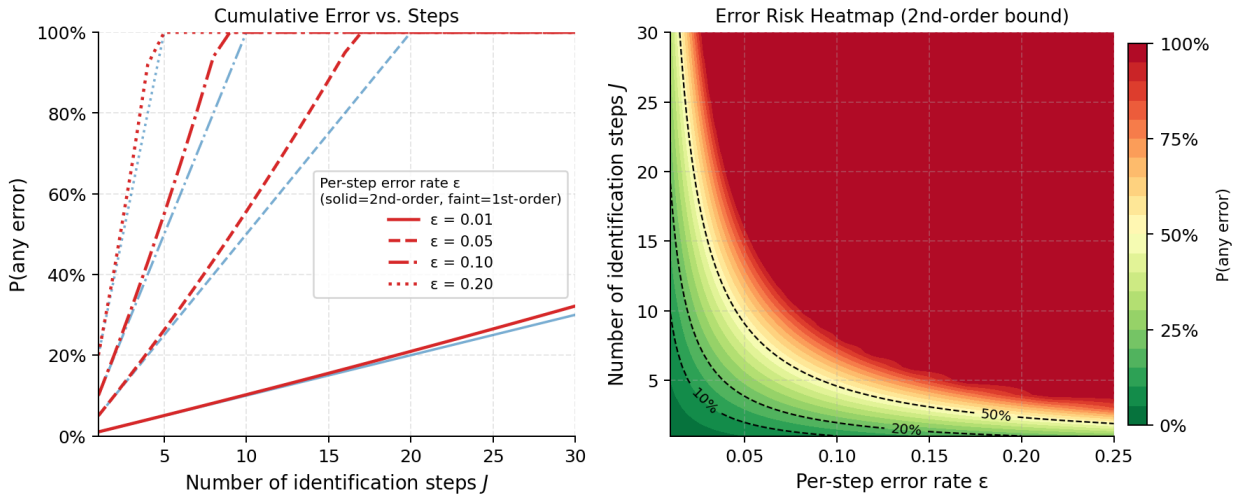

Figure 5: Error propagation in sequential CoT identification (Δ_prop = 0.5). Left: cumulative error probability vs. number of steps J for four per-step error rates ε; faint lines show the first-order union bound Jε, solid lines show the second-order bound including propagation (Proposition 5.1 part iii). Right: heatmap of 2nd-order cumulative error over the (ε, J) plane — contours at 10%, 20%, and 50% mark the practical safety boundaries.

5.7 Summary of Theoretical Results

This section established the four core theoretical results:

Theorem 6.2: Token lower bound $T = \Omega(J)$ — necessary tokens scale linearly with elements

Theorem 6.3: Success probability $P = (1 - e^{-\alpha T/J})^J$ — smooth scaling enables budget planning

Theorem 6.4: Threshold shift $J^*_{CoT}(T) \approx J^* + T$ — linear improvement in complexity frontier

Theorem 6.5: CoT-Fine-tuning equivalence for $J \leq J^*_{CoT}(T)$ — universal substitute without training

Together with Theorem 6.1 (Depth Amplification) from Section 3, these five theorems provide a complete theoretical framework for understanding and deploying chain-of-thought prompting in GeoAI. Section 6 translates these results into practical deployment guidelines and testable predictions.

6. Optimal Strategies and Guidelines

This section bridges theory and practice, synthesizing the theoretical results from Sections 3–5 into actionable deployment guidelines. We formalize strategy selection criteria, establish universal deployment guidelines, and derive eight testable predictions for empirical validation.

6.1 Strategy Selection Framework

Section 4 introduced domain-specific CoT strategies. Here we provide a systematic framework for selecting the optimal strategy based on task characteristics.

Corollary 6.2 (Strategy Selection Criteria). For J-element detection, select strategy based on:
(i) Spatial decomposition: When elements are distributed across a 2D/3D spatial domain with approximate locations determinable by region. Optimal for remote sensing and multi-modal VQA.
(ii) Temporal scanning: When elements evolve over time or have temporal dependencies. Essential for climate event detection and change analysis.
(iii) Confidence-ordered: When element saliency varies significantly and high-confidence detections should precede uncertain ones. Preferred for geodesy and sparse detection.
(iv) Hierarchical: When J is large ($J > 30$) and coarse-to-fine refinement reduces per-level complexity below J^*.
(v) Verification loops: When accuracy is critical (safety applications) and the 20–30% token overhead for verification is acceptable.

Table 8 summarized the strategy selection by domain and task characteristics.

Table 8: Strategy selection matrix by domain and task characteristics

Domain	Primary	Secondary	When to Add Verification
Remote Sensing	Spatial decomp.	Confidence-ord.	Infrastructure monitoring
Climate	Temporal scan	Hierarchical	Early warning systems
Geodesy	Confidence-ord.	Spatial decomp.	Seismic/hazard analysis
Multi-Modal	Region enumeration	Verification	Navigation, accessibility

6.2 Universal Deployment Guidelines

We synthesize the theoretical results into a step-by-step deployment framework applicable across all GeoAI domains.

Corollary 6.3 (Universal Deployment Guidelines). For J-element detection in any GeoAI domain:
Step 1 — Estimate J: If J is unknown, use light CoT ($T \approx 20$) for initial estimate, then refine allocation.
Step 2 — Select approach based on J:

Table 9: Approach selection based on estimated J

Estimated J	Approach	Token Budget	Strategy
$J \leq 4$	Standard ICL	$T = 0$	Direct prediction
$4 < J \leq 10$	Light CoT	$T = 15–25$	Single-pass enum.
$10 < J \leq 25$	Moderate CoT	$T = 30–60$	Spatial/temporal
$25 < J \leq 50$	Heavy CoT	$T = 70–120$	Hierarchical + verif.
$J > 50$	Fine-tuning or extended	$T > 120$ or FT	Task-specific training

Step 3 — Apply domain multiplier: Compute precise budget as $T = \kappa_d \cdot J + C_d$ using domain-specific constants from Table 5.
Step 4 — Add verification for high-stakes: For safety-critical applications, increase budget by 25% and include explicit verification phase.

6.3 Token Budget Calculator

For convenience, we provide precomputed token budgets for common scenarios.

Table 10: Recommended token budgets by domain and target J

Target J	RS	Multi-Modal	Geodesy	Climate	+Verif.
5	23	30	24	33	+25%
10	35	45	41	55	+25%
20	60	75	76	100	+25%
30	85	105	111	145	+25%
50	135	165	181	235	+25%

Values computed using $T = \kappa_d \cdot J + C_d$ with κ_RS = 2.5, κ_mm = 3.0, κ_geo = 3.5, κ_clim = 4.5 and corresponding overhead constants. The "+Verif." column indicates the additional budget for verification-enabled strategies.

6.4 Testable Predictions

The theoretical framework yields eight quantitative predictions that enable empirical validation. Each prediction includes a testable form and suggested experimental design.

Prediction P1 (Token-Accuracy Scaling Law). Detection accuracy follows:
$Accuracy(T, J) = (1 - exp(-\alpha T/J))^J$
Testable form: Plot accuracy vs. T/J across multiple (T, J) combinations. All curves should collapse to a single master curve when parameterized by T/J. Validate by fitting α and measuring R^2 of fit.

Prediction P2 (Threshold Shift Validation). The practical detection limit shifts linearly:
$J^*_{CoT}(T) \approx 3 + T$
Testable form: For each $T \in \{0, 10, 20, 30, 50, 100\}$, find J^* where accuracy drops below 80%. Plot J^* vs. T; verify linear relationship with slope ≈ 1, intercept ≈ 3.

Prediction P3 (Spatial Decomposition Superiority). For $J > 10$:
$Accuracy_{spatial}(T, J) - Accuracy_{holistic}(T, J) \in [0.15, 0.25]$
Testable form: Compare quadrant-based CoT vs. single-pass CoT with equal token budget. Measure accuracy difference across $J \in \{12, 15, 20, 25, 30\}$; verify 15–25% improvement.

Prediction P4 (Domain Invariance). Accuracy curves align when normalized by domain multiplier:

Accuracy_d(T, J) ≈ Accuracy_{universal}(T/κ_d, J)
Testable form: Collect *(T, J, Accuracy)* data from each domain. Rescale $T \rightarrow T/\kappa_d$; verify curves collapse to universal form (RMSE < 0.05).

Prediction P5 (Verification Benefit). Verification improves accuracy at known cost:
$\Delta Accuracy \in [0.10, 0.15], \quad \Delta T/T_{base} \in [0.20, 0.30]$
Testable form: Run same task with and without verification phase. Measure accuracy improvement and token overhead; compute efficiency ratio *ΔAccuracy/ΔTokens*.

Prediction P6 (Diminishing Returns). Marginal accuracy improvement decreases:
$\partial Accuracy/\partial T \propto 1/T \quad$ *for large T*
Testable form: Measure accuracy at $T \in \{20, 30, 40, 50, 60, 80, 100, 150\}$. Compute numerical derivative; log-log plot of $\partial Accuracy/\partial T$ vs. *T* should have slope ≈ −1.

Prediction P7 (Error Propagation). Early errors increase subsequent error probability:
$P(\text{error at } t \mid \text{error at } t-1) > P(\text{error at } t)$
Testable form: Inject controlled errors at step $t-1$; measure error rate at step *t* with vs. without injection. Estimate propagation penalty Δ_{prop}; verify $\Delta_{prop} > 0$.

Prediction P8 (Fine-Tuning Crossover). CoT efficiency crosses fine-tuning at $J_{cross} \approx 50$:
$C_{CoT}(J) > C_{FT}(J) \quad$ *for* $J > J_{cross}$
Testable form: Measure total compute (training + inference) for CoT at various *J* vs. fine-tuned models. Find crossover point; verify $J_{cross} \in [40, 60]$.

6.5 Summary

This section translated the theoretical framework into practical deployment tools:

- **Strategy selection criteria** (Corollary 6.2): Domain-specific strategy recommendations
- **Universal deployment guidelines** (Corollary 6.3): Four-step framework applicable across domains
- **Token budget calculator** (Table 10): Precomputed budgets for rapid deployment
- **Testable predictions** (P1–P8): Quantitative hypotheses for empirical validation

The eight predictions provide a comprehensive test suite for validating the theoretical framework empirically. Successful validation would establish CoT depth amplification as a verified mechanism, while failures would identify areas for theoretical refinement.

7. Discussion

This section situates our results within the broader context of foundation model research, discusses implications for GeoAI practice, acknowledges limitations of the theoretical framework, and identifies directions for future investigation.

7.1 The Theoretical Arc

This paper plays a critical role in the theoretical program that begun in Paper 1 (Hawarey 2026a); it's the 6th paper out of 7. That work introduced the ICL characterization framework based on Sufficient Statistic Complexity (SSC) and established the fundamental ICL-Easy/Hard dichotomy. Papers 2–5 (Hawarey 2026b-e) systematically mapped this dichotomy across GeoAI domains—geodesy, remote sensing, climate science, and multi-modal applications—establishing that the hardness threshold $J^* \approx 2–4$ is remarkably consistent across domains despite their diverse physical foundations.

The present paper provides the resolution: chain-of-thought prompting is a *universal mechanism* for overcoming ICL-Hard barriers. The key insight—*that autoregressive token generation provides computational depth amplification*—explains both *why* CoT works and *how much* it helps. The threshold shift formula $J^*_{CoT}(T) \approx J^* + T$ quantifies this improvement precisely.

The theoretical arc thus has a satisfying structure: Papers 1–5 (Hawarey, 2026a-e) characterized the problem (what is hard and why), and Paper 6 (this one) provides the solution (how to overcome hardness). Together, with a planned 7th paper about (In-Context Learning Characterization of Navigation and GNSS Foundation Models), the seven papers constitute a comprehensive theory of in-context learning for geospatial AI, with practical implications for foundation model deployment.

7.2 Implications for Foundation Model Design

Our results have several implications for the design and training of geospatial foundation models.

CoT-aware training. Current GeoFMs are typically trained for single-pass prediction. Our results suggest that incorporating CoT traces during training—either through explicit supervision or reinforcement learning from intermediate reasoning—could improve performance on ICL-Hard tasks without increasing model depth. This represents a training-time investment that pays off at inference time through more efficient CoT generation.

Adaptive computation. The framework suggests a natural adaptive computation strategy: estimate J from initial context, allocate tokens according to $T = \kappa_d \cdot J + C_d$, and generate CoT accordingly. This enables task-appropriate depth without architectural changes—simple tasks use few tokens while complex tasks use many.

Depth vs. width tradeoffs. The equivalence between token generation and depth amplification suggests new tradeoffs in model design. Rather than building deeper models (which are harder to train and slower to run), practitioners might prefer shallower models that can generate more tokens when needed. The compute tradeoff is favorable when task complexity varies widely.

Multi-modal integration. For vision-language GeoFMs like GeoChat, our results suggest that effective multi-modal CoT requires the vision encoder to provide information that can be progressively refined through language generation. This argues for tighter integration between vision and language components rather than simple feature concatenation.

7.3 Connections to Related Work

Our theoretical framework connects to several lines of research in machine learning and complexity theory.

Chain-of-thought prompting. Wei et al. (2022) introduced CoT prompting and demonstrated dramatic improvements on arithmetic and reasoning tasks. Subsequent work explored variations including zero-shot CoT (Kojima et al. 2022), self-consistency (Wang et al. 2023), and tree-of-thought (Yao et al. 2023). Our work provides theoretical grounding for these empirical observations: CoT succeeds because it provides computational depth that single-pass inference lacks.

Transformer expressiveness. Theoretical work on transformer capabilities has established connections to circuit complexity (Merrill & Sabharwal 2023), shown limitations on certain computations (Hahn 2020), and characterized the role of depth (Sanford et al. 2023). Foundational results on transformer computational power include Weiss et al. (2021), who introduced the RASP programming language to characterize what transformers can compute, and Pérez et al. (2021), who proved that transformers with hard attention are Turing complete under certain conditions—results that sharpen the boundary between computable and incomputable functions in this model class. Particularly relevant to the present work are Feng et al. (2023), who formally connected CoT to circuit complexity, and Li et al. (2024), who proved that CoT empowers transformers to solve inherently serial problems. Our analysis extends this line by showing how autoregressive generation escapes depth limitations in the specific context of GeoAI, providing domain-specific token complexity bounds.

In-context learning theory. Recent theoretical work has analyzed ICL through various lenses: as implicit Bayesian inference (Xie et al. 2022), as gradient descent (von Oswald et al. 2023), and as algorithm learning (Garg et al. 2022). Our SSC/AC-SSC framework complements these perspectives by providing a complexity-theoretic characterization that directly connects to circuit bounds.

Geospatial foundation models. The rapid development of GeoFMs including SatMAE, Scale-MAE, Prithvi, and Pangu-Weather has outpaced theoretical understanding. Our framework provides principled guidance for when these models will succeed (ICL-Easy tasks) versus struggle (ICL-Hard tasks), and how to extend their capabilities through CoT.

7.4 Limitations

We acknowledge several limitations of the theoretical framework that suggest directions for refinement.

Absence of empirical validation. This is the primary limitation of the present work: the entire theoretical framework—including the depth amplification formula, the token complexity bounds, the success probability scaling law, and the domain-specific multipliers—has not been empirically validated. No experimental results are reported, and none of the eight testable predictions (P1–P8 in Section 6.4) have been verified against real GeoFM outputs. The

framework is therefore a theoretical proposal awaiting empirical confirmation. Practitioners should treat the deployment guidelines (Section 6) as theoretically motivated hypotheses rather than validated prescriptions. Empirical validation is the most important direction for future work.

Idealized model. The analysis assumes idealized conditions: perfect token generation, no context length limitations, and uniform token allocation. Real systems face context window constraints, imperfect generation, and uneven difficulty across elements. The practical token multipliers κ_d partially absorb these effects but may require empirical calibration.

Independence assumption. Theorem 6.3's success probability derivation assumes independence across element detections. In practice, elements may cluster spatially or have correlated saliency, violating independence. The error propagation analysis (Propositions 5.1–5.2) partially addresses this but a full treatment of correlated detection remains open.

Domain multiplier estimation. The domain-specific multipliers $\kappa_d \in [2.5, 4.5]$ are derived from first-principles decomposition but require empirical validation. The true multipliers may vary with model architecture, training data, and prompt design. The testable predictions (Section 6.4) provide a validation pathway.

Worst-case vs. average-case. The circuit complexity reductions establish worst-case hardness, but practical performance depends on average-case behavior over natural input distributions. GeoAI tasks may have favorable structure (e.g., spatially coherent objects) that makes them easier than worst-case bounds suggest.

Single-model focus. The analysis focuses on single-model CoT. Ensemble methods, model cascades, and retrieval-augmented generation may provide alternative or complementary mechanisms for overcoming ICL-Hard barriers. A unified treatment of these approaches remains for future work.

Foundational dependency on prior series papers. The present paper depends foundationally on the results established in Hawarey (2026a–e). The ICL-Hard characterization, the SSC/AC-SSC framework, and the domain-specific hardness thresholds are all taken as given from those companion papers. If any of those foundational results were found to be in error—for example, if the reduction to sparse parity were shown to be invalid for a particular domain, or if the hardness threshold J* were substantially different from O(log log n) in practice—the conclusions of this paper would be affected accordingly. This series-level dependency represents a structural risk for the theoretical program as a whole.

7.5 Open Problems

The theoretical framework suggests several open problems for future investigation.

Open Problem 1 (Optimal CoT Structure). What is the optimal structure of CoT traces for GeoAI tasks? Is there a learnable policy that allocates tokens adaptively based on intermediate results? Can reinforcement learning discover more efficient CoT strategies than hand-designed decompositions?

Open Problem 2 (Continuous vs. Discrete Elements). Our framework focuses on discrete element identification. How does the analysis extend to continuous-valued outputs (e.g., precise localization, intensity estimation)? Is there a continuous analog of the hardness threshold?

Open Problem 3 (Multi-Scale Detection). GeoAI tasks often involve multi-scale phenomena (e.g., regional vs. local features). How should CoT strategies handle hierarchical structure? Is there an optimal scale decomposition that minimizes total tokens?

Open Problem 4 (Uncertainty Quantification). The current framework focuses on point estimates. How should CoT be designed to provide calibrated uncertainty estimates for detected elements? Can verification phases provide uncertainty quantification without separate uncertainty models?

7.6 Broader Impact

The practical deployment of CoT for GeoAI applications carries both opportunities and responsibilities.

Positive applications. Improved object detection enables better disaster response (locating survivors, assessing damage), environmental monitoring (tracking deforestation, detecting pollution), and urban planning (infrastructure assessment, traffic management). Climate event detection supports early warning systems that save lives. Multi-modal GeoAI enhances accessibility of geospatial information to non-experts.

Potential misuse. The same capabilities enable surveillance applications, military targeting, and privacy violations. Improved detection of vehicles, individuals, or facilities from satellite imagery raises significant ethical concerns. We encourage responsible deployment with appropriate oversight, particularly for dual-use applications.

Verification and trust. CoT provides a form of interpretability—the reasoning trace can be inspected to understand model decisions. This transparency is valuable for high-stakes applications where users need to verify model outputs. However, CoT traces may not faithfully represent internal computations, and over-reliance on apparent reasoning could be misleading.

Computational costs. CoT increases inference compute linearly with token count. For large-scale deployments processing millions of images or climate records, this overhead is substantial. We encourage research into more efficient CoT mechanisms and careful cost-benefit analysis for deployment decisions.

8. Conclusions

This paper has established chain-of-thought prompting as a universal mechanism for overcoming ICL-Hard barriers in geospatial artificial intelligence. By formalizing CoT as computational depth amplification, we have provided both theoretical foundations and practical deployment guidelines for extending foundation model capabilities across GeoAI domains.

Our main contributions are:

1. **Depth Amplification Theory (Theorem 6.1).** We proved that generating T chain-of-thought tokens provides effective computational depth $L_{eff}(T) = L + \gamma T$, enabling transformers to escape AC^0/TC^0 circuit limitations. This formalizes why CoT works: it provides the sequential computation that single-pass inference lacks.
2. **Tight Token Bounds (Theorems 6.2–6.3).** We established that $T = \Omega(J)$ tokens are necessary and $T = O(J \log J)$ suffice for J-element detection. The success probability scales as $P = (1 - e^{-\alpha T/J})^J$, enabling principled token budget allocation.
3. **Threshold Shift Quantification (Theorem 6.4).** The effective hardness threshold shifts from $J^* \approx 2–4$ to $J^*_{CoT}(T) \approx J^* + T$. With 100 tokens, tasks involving over 100 discrete elements become tractable—a 34× improvement over standard ICL.
4. **Domain-Specific Strategies (Section 4).** We derived CoT strategies for all four GeoAI domains: spatial decomposition for remote sensing ($\kappa_{RS} \approx 2.5$), region enumeration for multi-modal ($\kappa_{mm} \approx 3.0$), sequential search for geodesy ($\kappa_{geo} \approx 3.5$), and temporal scanning for climate ($\kappa_{clim} \approx 4.5$).
5. **Practical Guidelines (Section 6).** We synthesized the theory into actionable deployment guidelines: strategy selection criteria, token budget calculators, and eight testable predictions for empirical validation.

Together with the companion Papers 1–5 and the planned 7th. paper, this work provides a theoretical framework for understanding in-context learning in geospatial AI. The framework explains the observed dichotomy between tasks where foundation models excel (ICL-Easy: classification, segmentation, smooth field prediction) and tasks where they struggle (ICL-Hard: object detection, source localization, event identification, counting). More importantly, it provides a principled solution through chain-of-thought prompting, with quantitative guidance on token allocation and strategy selection.

The practical impact is substantial. Foundation models like SatMAE, Prithvi, Pangu-Weather, and GeoChat can now be deployed on ICL-Hard tasks that previously required task-specific fine-tuning, simply by enabling CoT generation with appropriate token budgets. This democratizes access to advanced GeoAI capabilities: users without training infrastructure or domain-specific datasets can leverage foundation models for complex detection and counting tasks through prompt engineering alone.

Looking forward, we envision several directions for extending this work. A planned 7th paper about (In-Context Learning Characterization of Navigation and GNSS Foundation Models) would complete a 7-paper research series to constitute a comprehensive theory of in-context learning for geospatial AI, with practical implications for foundation model deployment. Also, empirical validation of the eight predictions (Section 6.4) will refine the theoretical parameters and identify areas for improvement. CoT-aware training could further improve efficiency by teaching models to generate more effective reasoning traces. Integration with retrieval-augmented generation and model cascades may provide complementary mechanisms for complex tasks. Finally, extension to streaming and real-time applications—critical for climate early warning and disaster response—presents important engineering challenges.

In conclusion, chain-of-thought prompting transforms the landscape of what geospatial foundation models can achieve. By providing computational depth through token generation, CoT breaks through the $J^* \approx 2–4$ barrier that fundamentally limited standard in-context learning. The result is a unified theory connecting circuit complexity, transformer computation, and practical GeoAI deployment—a theory that both explains existing observations and guides future development of Earth observation AI systems.

9. Acknowledgment

This is an Artificial Intelligence-Assisted Research (AIAR). AI tools were employed as efficiency-enhancing assistants. The author declares no conflict of interest. This study received no external funding.

10. References

1. Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature, 619*, 533–538.
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.
3. Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., ... & Ni, B. (2022). SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems, 35*, 197–211.
4. Feng, G., Zhang, B., Gu, Y., Ye, H., He, D., & Wang, L. (2023). Towards revealing the mystery behind chain of thought: A theoretical perspective. *Advances in Neural Information Processing Systems, 36*, 70757–70798.
5. Garg, S., Tsipras, D., Liang, P., & Valiant, G. (2022). What can transformers learn in-context? A case study of simple function classes. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 30583–30598.
6. Hahn, M. (2020). Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics, 8*, 156–171.
7. Håstad, J. (1987). Computational limitations of small-depth circuits. *MIT Press*.
8. Hawarey, M. (2026a). Characterizing In-Context Learning: When Can Transformers Match Standard Learning Algorithms? AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026277, DOI: 10.65737/AIRMCS2026277.
9. Hawarey, M. (2026b). Quantum-Enhanced In-Context Learning for Geopotential Field Estimation: A Theoretical Framework. AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026322, DOI: 10.65737/AIRMCS2026322.
10. Hawarey, M. (2026c). ICL Characterization of Geo-Foundation Models. AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026360, DOI: 10.65737/AIRMCS2026360.
11. Hawarey, M. (2026d). ICL Characterization of Climate Foundation Models: When Can Transformers Learn Weather and Climate? AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026405, DOI: 10.65737/AIRMCS2026405.

12. Hawarey, M. (2026e). ICL Characterization of Multi-Modal Geo-Foundation Models: When Can Vision-Language Transformers Learn Geospatial Tasks? AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026446, DOI: 10.65737/AIRMCS2026446.
13. Jakubik, J., Roy, S., Phillips, C. E., Fraccaro, P., Godwin, D., Zadrozny, B., ... & Mukkavilli, K. (2023). Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*.
14. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems, 35*, 22199–22213.
15. Kuckreja, K., Danish, M. S., Naseer, M., Das, A., Khan, S., & Khan, F. S. (2024). GeoChat: Grounded large vision-language model for remote sensing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27831–27840.
16. Li, Z., Huang, B., Huang, H., Lu, P., Wu, Z., & Zhang, L. (2024). Chain of thought empowers transformers to solve inherently serial problems. *Proceedings of the 12th International Conference on Learning Representations*.
17. Merrill, W., & Sabharwal, A. (2023). The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics (TACL)*, 11, 531–545.
18. Pérez, J., Barceló, P., & Marinkovic, J. (2021). Attention is Turing complete. *Journal of Machine Learning Research*, 22(75), 1–35.
19. Reed, C. J., Guber, R., Wang, S., Malik, A., Ahlgren, O., Keutzer, K., & Darrell, T. (2023). Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4088–4099.
20. Sanford, C., Hsu, D., & Telgarsky, M. (2023). Representational strengths and limitations of transformers. *Advances in Neural Information Processing Systems, 36*, 38693–38716.
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.
22. von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., & Vladymyrov, M. (2023). Transformers learn in-context by gradient descent. *Proceedings of the 40th International Conference on Machine Learning*, 35151–35174.
23. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. *Proceedings of the 11th International Conference on Learning Representations*.
24. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 24824–24837.
25. Weiss, G., Goldberg, Y., & Yahav, E. (2021). Thinking like transformers. *Proceedings of the 38th International Conference on Machine Learning*, 139, 11080–11090.
26. Xie, S. M., Raghunathan, A., Liang, P., & Ma, T. (2022). An explanation of in-context learning as implicit Bayesian inference. *Proceedings of the 10th International Conference on Learning Representations*.
27. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems, 36*, 11809–11822.