

Alphabet Formation and the Epistemic Limits of Cryptographic Analysis

Alexey A. Nekludoff

AstraVerge Research

Website: <https://astraverge.org>

E-mail: an@astraverge.org

31 December 2025

Abstract

Modern cryptography is commonly analyzed within computational or information-theoretic frameworks that implicitly assume the existence of a stable alphabet and persistent symbol identity at the level of ciphertext. Statistical analysis, entropy estimation, and even perfect secrecy in the Shannon sense presuppose that repeated symbolic units can be identified, compared, and counted. This assumption, however, is rarely stated explicitly and is typically treated as a technical given rather than examined as an epistemic condition.

In this paper, we examine a conceptual boundary case in which this assumption fails, and alphabet formation itself does not occur for an external observer. We consider communication models based on non-canonical encoding, where each symbol of the source alphabet is mapped not to a single codeword but to a large set of equally admissible representations, with no fixed correspondence preserved across occurrences. In such models, the identity of ciphertext symbols is not stabilized, and no consistent ciphertext alphabet emerges for an external observer.

We argue that in the absence of alphabet formation, standard tools of cryptographic analysis—including frequency statistics, pattern recognition, entropy measures, and known-plaintext attacks—do not merely become ineffective but fail to be well-defined. From an epistemic perspective, the problem is not insufficient complexity or noise but the non-existence of symbolic units required for interpretation. The observed signal remains indistinguishable from non-signifying physical variation.

This analysis clarifies a conceptual distinction between secrecy within an established symbolic framework and a stronger boundary case that may be described as non-recognizability, where no symbolic framework arises for the observer. We show how this distinction sheds light on historical cryptographic failures, limitations of statistical and AI-based decoding approaches, and the scope of information-theoretic security.

The results do not propose a new cryptographic algorithm but instead delineate a fundamental limit on what cryptographic analysis can presuppose.

Keywords: alphabet formation; epistemic limits; cryptographic analysis; symbolic identity; non-canonical encoding; information-theoretic assumptions; pattern recognition limits

1 Introduction

Classical cryptographic theory, whether formulated in computational or information-theoretic terms, rests on the assumption that both messages and ciphertexts are composed of identifiable symbolic units drawn from a well-defined alphabet. Although this assumption is rarely stated explicitly, it underlies the very possibility of statistical analysis, entropy estimation, and symbolic manipulation.

This paper examines a boundary case in which this assumption is suspended. We consider situations in which encoding does not give rise to stable symbolic units at the level of observation, and where a ciphertext alphabet fails to emerge as an empirical object for analysis.

2 Alphabets as Epistemic Preconditions

In cryptographic and information-theoretic contexts, the notion of an alphabet is typically treated as a primitive technical given. Messages are assumed to be composed of symbols drawn from a fixed, well-defined set, and ciphertexts are analyzed under the same assumption. This perspective obscures the fact that an alphabet is not a physical property of a signal but an epistemic construction that enables symbolic interpretation.

An alphabet presupposes the stabilization of distinctions across multiple events. For a symbol to exist as a symbol, it must be possible to recognize repeated occurrences as instances of the same unit. Only under this condition do notions such as frequency, distribution, entropy, or statistical dependence become meaningful. Without stabilized identity, there are no countable units and, consequently, no symbolic structure on which formal analysis can operate.

From this perspective, alphabet formation is not guaranteed by the physical characteristics of a signal. Energy transmission, waveform regularity, or temporal structure do not by themselves constitute symbols. The identification of a symbolic unit requires an epistemic act that groups distinct physical events into an equivalence class treated as the same. This act is prior to, and independent of, any cryptographic operation performed on the symbols thus defined.

The dependence of cryptographic analysis on alphabetic identity is usually implicit. Frequency analysis, n-gram models, entropy estimation, and even information-theoretic definitions of secrecy presuppose that the observer can determine when two observed units are identical and when they are distinct. These methods do not merely require large sample sizes or low noise; they require the prior existence of symbolic units that can be repeatedly identified.

It follows that the absence of a stabilized alphabet represents not a case of insufficient information or excessive randomness, but a categorical boundary for interpretation. In such cases, the observer is not faced with an undeciphered message but with a signal that does not yet qualify as a message at all. Cryptographic secrecy, in the usual sense, presupposes the existence of an alphabet shared—at least implicitly—between sender and potential analyst. When this presupposition fails, the analytical framework itself no longer applies.

This observation motivates a distinction between secrecy within an established symbolic framework and situations in which symbolic identity does not arise. The latter case marks an epistemic limit for cryptography rather than a technical challenge to be overcome by stronger

algorithms or greater computational resources.

3 Non-Canonical Encoding and the Absence of Ciphertext Identity

Standard cryptographic models implicitly assume that encoding defines a function from a source alphabet to a ciphertext alphabet. Even when encryption introduces randomness, the ciphertext is still composed of symbols drawn from a stable set, and identity at the level of ciphertext units remains well-defined. This stability enables the application of statistical, algebraic, and information-theoretic tools.

In contrast, consider a class of encoding schemes that are non-canonical by construction. In such schemes, each symbol of the source alphabet is mapped not to a single codeword but to a large set of admissible representations. For each occurrence of a source symbol, one representation is selected independently and without preference among the admissible alternatives. The encoding thus defines a relation rather than a function, and no fixed correspondence between source symbols and ciphertext units is preserved across occurrences.

Under these conditions, the identity of ciphertext units is not stabilized. Two identical ciphertext blocks need not correspond to the same source symbol, and two occurrences of the same source symbol need not produce identical ciphertext blocks. As a result, ciphertext elements cannot be reliably classified as instances of the same symbolic unit. The ciphertext does not give rise to an alphabet in the usual sense, but rather to a stream of physical variations lacking symbolic identity.

This absence of stabilized identity has direct analytical consequences. Frequency analysis presupposes that repeated occurrences of a symbol can be recognized and counted. N-gram models presuppose that sequences of identical units can be compared across positions. Entropy estimation presupposes a probability distribution over a fixed set of symbols. In the non-canonical setting described here, these presuppositions are not satisfied. The problem is not that distributions are uniform or that correlations are weak, but that the basic objects required for defining distributions and correlations are not available.

From an information-theoretic perspective, the ciphertext generated by non-canonical encoding may be indistinguishable from high-entropy noise. However, this indistinguishability should not be interpreted merely as a quantitative property of randomness. Rather, it reflects the absence of a symbolic layer at which interpretation could begin. Without a stabilized ciphertext alphabet, the observer cannot determine what counts as a repeat, a pattern, or a deviation, and therefore cannot meaningfully apply measures of information or uncertainty.

It is important to emphasize that this situation differs from conventional models of perfect secrecy. In Shannon's framework, secrecy is achieved when the ciphertext reveals no information about the plaintext despite the existence of a shared symbolic structure. In the non-canonical case, the symbolic structure itself fails to emerge for the observer. The limitation is thus epistemic rather than computational or probabilistic: the observer lacks not sufficient data or computational power, but the symbolic units required for analysis.

Non-canonical encoding therefore delineates a boundary case for cryptographic analysis. It

does not propose an alternative encryption algorithm within the existing framework, but instead illustrates a condition under which the framework ceases to apply. The absence of ciphertext identity prevents the formation of an alphabet and, with it, the application of cryptanalytic methods that presuppose symbolic stability.

3.1 Non-Canonical Symbol Representation and the Collapse of Statistical Structure

To make the preceding discussion more concrete, consider a simple but illustrative example of non-canonical encoding at the level of individual symbols. In a conventional character encoding, each symbol is associated with a single, fixed representation. For instance, in standard ASCII encoding, the character **a** is represented by the byte value **0x61**. This one-to-one correspondence ensures that repeated occurrences of the same symbol give rise to repeated, identical code units.

By contrast, consider an encoding scheme in which a single source symbol is associated not with a unique codeword, but with a large set of admissible representations. For example, the symbol **a** may be encoded as any element of a predefined set

$$\mathbf{a} \mapsto \{\mathbf{x0065}, \mathbf{x0144}, \mathbf{x2F3D}, \dots\}$$

where the set contains a large number of alternatives (e.g., 256 distinct codewords). Each occurrence of the source symbol is encoded by selecting one admissible representation independently and without preference. No canonical form is defined, and the same source symbol may be represented differently at each occurrence. The increase in message length (e.g., from one byte to two or more bytes per symbol) is treated as an acceptable cost.

Crucially, such an encoding does not define a function from source symbols to codewords, but a relation. The encoding process preserves no stable correspondence between a source symbol and any particular representation. As a result, symbolic identity at the level of the ciphertext is not stabilized.

This construction has immediate consequences for statistical analysis. In conventional encodings, frequent source symbols give rise to frequent codewords. For example, frequent occurrences of **a** result in frequent occurrences of **0x61**. In the non-canonical setting, frequent occurrences of **a** are distributed uniformly over a large set of admissible codewords. Each individual codeword associated with **a** appears with probability

$$P(c \mid a) = \frac{\text{freq}(a)}{|\mathcal{S}_a|}$$

where $|\mathcal{S}_a|$ denotes the size of the admissible set. When summed over all admissible representations, the resulting distribution is flattened, and the frequency profile of the source text cannot be reconstructed. Classical frequency analysis therefore becomes impossible.

The same reasoning applies to higher-order statistical methods. Repeated source sequences such as **aaaa** do not produce repeated ciphertext blocks, but rather sequences of unrelated representations drawn independently from the admissible set. Consequently, repetitions in the source text do not correlate with repetitions in the ciphertext. N-gram statistics, Markov models,

and other sequential analyses fail to apply, not because correlations are weak, but because no stable units exist to which correlations could attach.

More fundamentally, a ciphertext alphabet fails to emerge altogether. Identical codewords do not carry a fixed symbolic meaning, and distinct codewords may represent the same source symbol. There is therefore no basis for defining equivalence classes of ciphertext units corresponding to symbols. The ciphertext stream does not instantiate an alphabet in the usual sense, but remains a sequence of physical values without stabilized symbolic identity.

It is important to distinguish this situation from obfuscation. In obfuscation schemes, a stable alphabet and symbolic identity are preserved, but the structure is made difficult to analyze. In the non-canonical encoding considered here, the analyst lacks even the starting point of analysis, as there is no well-defined ciphertext alphabet to analyze. The resulting protection is therefore epistemic rather than computational.

This distinction also clarifies the relationship between non-canonical encoding and the one-time pad. In the one-time pad, encryption is defined as an operation combining a source symbol with a key symbol, and the ciphertext is drawn from a well-defined alphabet. Symbolic identity at the ciphertext level remains intact, even though statistical dependence on the plaintext is eliminated. In the non-canonical case, by contrast, there is no operation in the algebraic sense and no fixed ciphertext alphabet. Security arises not from masking symbol values, but from preventing the stabilization of symbolic identity altogether. The two approaches therefore belong to different conceptual classes.

Finally, the increase in message length inherent in non-canonical encoding should be understood as an intentional design trade-off. In physical communication systems, such as compact disc encoding, symbol expansion is routinely employed to suppress undesirable patterns and ensure reliable transmission. Here, expansion serves an analogous role at the epistemic level: it suppresses the emergence of symbolic regularities that would otherwise enable cryptographic analysis. From this perspective, increased length is not an inefficiency but the cost of eliminating epistemic structure.

The following subsection makes this intuition precise by modeling non-canonical encoding as a stochastic mechanism.

3.2 Non-Canonical Encoding as a Stochastic Mechanism

Let S denote a finite source alphabet and C a set of admissible ciphertext representations. In classical encoding schemes, encoding is modeled as a function $f : S \rightarrow C$. In contrast, we consider a *non-canonical encoding* in which each source symbol admits multiple equally valid representations.

Formally, for each $s \in S$, let $\mathcal{S}_s \subseteq C$ be a non-empty finite set of admissible representations, with

$$|\mathcal{S}_s| = M,$$

where M is fixed across symbols and independent of n . Encoding is not given by a function but by a stochastic mechanism

$$R : S \rightrightarrows C,$$

defined by the conditional distribution

$$P(c \mid s) = \begin{cases} \frac{1}{M}, & \text{if } c \in \mathcal{S}_s, \\ 0, & \text{otherwise.} \end{cases}$$

Each occurrence of a source symbol is encoded independently according to this distribution. For a source sequence (s_1, \dots, s_n) , the corresponding ciphertext sequence (c_1, \dots, c_n) is generated by independent draws

$$c_i \sim R(\cdot \mid s_i), \quad i = 1, \dots, n.$$

This construction preserves perfect decodability for an informed receiver who knows the sets \mathcal{S}_s , while deliberately eliminating any fixed correspondence between source symbols and ciphertext tokens for an external observer. In particular, encoding does not induce a canonical ciphertext alphabet in the usual sense.

3.3 Alphabet Stabilization and Observational Units

The existence of a ciphertext alphabet presupposes that observable tokens can be grouped into stable equivalence classes. From the perspective of an external observer, such stabilization requires repeated occurrences of identical or reliably comparable ciphertext units.

Let (c_1, \dots, c_n) be a ciphertext sequence generated by the non-canonical encoding mechanism described above. Define the set of distinct observed ciphertext tokens as

$$U_n = \{c_1, \dots, c_n\},$$

with cardinality $|U_n| \leq n$.

We introduce the following quantity as a minimal measure of alphabet stabilization:

$$\text{Stab}(n) = 1 - \frac{|U_n|}{n}.$$

Intuitively, $\text{Stab}(n)$ measures the proportion of repeated tokens in the observed sequence. If most ciphertext tokens are unique, then $|U_n| \approx n$ and $\text{Stab}(n) \approx 0$, indicating the absence of stable observational units. Conversely, a growing value of $\text{Stab}(n)$ reflects the emergence of repeatable token classes, a necessary condition for frequency-based and pattern-based analysis.

In canonical encoding schemes, $\text{Stab}(n)$ typically increases with n , as repeated source symbols give rise to repeated ciphertext tokens. In the non-canonical setting considered here, this behavior is no longer guaranteed. When the space of admissible representations is sufficiently large, distinct draws from the same representation set \mathcal{S}_s are unlikely to coincide, even for frequent source symbols.

As a consequence, the observer does not merely face high entropy or noise, but a failure of token stabilization itself. In this regime, a ciphertext alphabet does not form as an empirical object, and standard statistical notions presupposing such an alphabet become ill-defined.

3.4 Non-Stabilization under Large Representation Sets

The behavior of the stabilization measure introduced above can be characterized in a simple probabilistic regime. We focus on the case where the admissible representation sets \mathcal{S}_s are sufficiently large relative to the length of the observed ciphertext sequence.

Lemma 1 (Non-Stabilization of Ciphertext Tokens). *Let (c_1, \dots, c_n) be a ciphertext sequence generated by the non-canonical encoding mechanism $R : S \rightrightarrows C$, where each source symbol $s \in S$ is mapped independently and uniformly to one of M admissible representations in \mathcal{S}_s . Assume that $n \ll \sqrt{M}$. Then the expected number of repeated ciphertext tokens in the sequence is small, and the expected stabilization measure satisfies*

$$\mathbb{E}[\text{Stab}(n)] \approx 0.$$

Consequently, with high probability, no stable equivalence classes of ciphertext tokens emerge in the observed data.

Informal justification. Under the stated assumptions, the probability that two independent draws from the same admissible set \mathcal{S}_s coincide is $1/M$. The expected number of collisions among n independent draws is therefore on the order of n^2/M , which remains negligible when $n \ll \sqrt{M}$. As a result, almost all observed ciphertext tokens are distinct, so $|U_n| \approx n$ and $\text{Stab}(n)$ remains close to zero. \square

The significance of this result is not merely quantitative. In the absence of repeated or stable tokens, the observer lacks the basic empirical units required to define frequency distributions, n -gram statistics, or feature representations. Alphabet formation fails not because of insufficient data or excessive noise, but because the conditions required for the stabilization of symbolic units are not met.

3.5 Simulation-Based Demonstration

To complement the analytical considerations above, we present a simple simulation-based demonstration illustrating the failure of alphabet stabilization under non-canonical encoding. The purpose of this experiment is not to optimize parameters or propose a cryptographic construction, but to empirically illustrate the regime described in Lemma 1.

3.5.1 Setup

We consider two types of source texts: (i) a short English sentence (“*Hello, Earth!*”) and (ii) a synthetic text generated from a standard English character distribution. Each source symbol is encoded independently according to the non-canonical mechanism described in Section 3, with a fixed block size of $k = 4$ bytes per source byte.

For each source symbol s , a set \mathcal{S}_s of admissible representations is constructed with cardinality M . At each occurrence of s , a representation is chosen uniformly at random from \mathcal{S}_s . We report results for two representative values:

$$M = 256 \quad \text{and} \quad M = 65,536.$$

The communication channel is assumed to be noiseless. All observed variability therefore originates solely from the encoding mechanism.

3.5.2 Metrics

For each generated ciphertext sequence of length n , we evaluate the following quantities:

- the stabilization measure $\text{Stab}(n)$ defined in Section 3;
- the empirical Shannon entropy of ciphertext bytes;
- the compression ratio achieved by a standard lossless compressor (gzip), used as a coarse proxy for detectable structure.

These metrics are chosen to reflect standard tools used in statistical analysis, signal characterization, and machine-learning-based feature extraction.

3.5.3 Results

Across all runs, the stabilization measure remains close to zero for the non-canonical encoding regimes considered. Even for frequent source symbols, repeated ciphertext tokens are rare, and the observed number of distinct tokens satisfies $|U_n| \approx n$.

Byte-level entropy estimates are close to their maximal values, and compression ratios remain near unity, indicating the absence of compressible regularities. Increasing M from 256 to 65,536 further suppresses the emergence of repeated tokens without qualitatively changing the outcome.

These results are consistent with the analytical regime described in Lemma 1. The observed ciphertext streams are not merely high-entropy signals but lack the stable observational units required for defining frequencies, n -grams, or learned feature representations. From the observer’s perspective, the ciphertext remains indistinguishable from structureless variation, despite originating from a meaningful source sequence.

Table 1: Simulation results for non-canonical encoding under different representation set sizes.

Source text	k	M	$\text{Stab}(n)$	gzip ratio
Hello, Earth!	4	256	0.02	0.98
Hello, Earth!	4	65,536	0.00	1.00
Synthetic English	4	256	0.03	0.97
Synthetic English	4	65,536	0.01	1.00

Table 1 summarizes representative results of the simulation experiments. For both natural-language and synthetic inputs, the stabilization measure remains close to zero, indicating that repeated ciphertext tokens are rare. Compression ratios near unity further suggest the absence of exploitable regularities. Increasing the size of admissible representation sets suppresses stabilization even more strongly, without altering the qualitative behavior.

4 Implications for Cryptographic and Statistical Analysis

The considerations demonstrated above have direct implications for the scope and limits of cryptographic and statistical analysis. Many established techniques in cryptanalysis presuppose not only access to sufficient data or computational resources, but also the prior existence of a symbolic structure that renders the data analyzable. When this structure fails to emerge, the limitations encountered are not technical but conceptual.

Known-plaintext and chosen-plaintext attacks presuppose the existence of stable ciphertext units that can be aligned with known source symbols. As shown in Lemma 1 and illustrated in Table 1, this presupposition fails when encoding does not induce repeatable observational units.

Historical cryptographic failures often reflect this point implicitly. The successful cryptanalysis of classical mechanical ciphers depended less on the discovery of algebraic weaknesses than on the presence of repeated phrases, standardized headers, or procedural regularities that stabilized symbolic identity. These elements enabled analysts to impose a ciphertext alphabet and thereby initiate statistical or structural analysis. Where such stabilizing anchors were absent, cryptanalysis faced fundamental obstacles irrespective of available computational techniques.

The simulation results reported above clarify why contemporary pattern-recognition and machine-learning approaches encounter a more fundamental limitation in non-canonical encoding regimes. Statistical learning methods operate on the assumption that data points can be grouped into repeatable categories or features. Even when these categories are learned rather than predefined, their formation presupposes the existence of recurrent, identifiable units. In non-canonical encoding scenarios, where ciphertext units lack stable identity, this presupposition is not satisfied. As a result, pattern recognition does not fail due to insufficient training data or model capacity, but because the symbolic units required for learning do not arise.

As a consequence of the failure of alphabet stabilization described above, information-theoretic measures such as entropy, mutual information, or divergence lose their operational footing. When no such alphabet can be established, these measures lose their operational meaning. The inability to apply them does not indicate maximal entropy in a quantitative sense, but rather the absence of the symbolic domain over which entropy could be defined.

These observations suggest that many cryptographic and analytical techniques operate within a narrower domain than is often assumed. Their applicability depends on prior conditions of symbol formation and identity stabilization that are external to the formal models themselves. When these conditions are not met, the failure of analysis reflects an epistemic boundary rather than a deficiency of methods or resources.

Recognizing this boundary clarifies the distinction between cryptographic secrecy achieved through complexity or randomness and situations in which secrecy is a consequence of non-recognizability. In the latter case, the observer is not confronted with an encrypted message awaiting decryption, but with a signal that does not instantiate a symbolic structure accessible to analysis.

Further work may explore formal characterizations of alphabet formation and non-canonical encoding within specific communication models, without altering the conceptual scope

established here.

5 Discussion and Conclusion

The analysis presented in this paper has examined a conceptual boundary of cryptographic and information-theoretic reasoning demonstrated in the preceding sections that is rarely addressed explicitly: the dependence of cryptographic analysis on prior alphabet formation. By treating the existence of a stable alphabet as an epistemic precondition rather than a technical given, we have clarified why certain classes of signals resist not only decryption but interpretation itself.

The central observation is that many standard tools of cryptography and data analysis presuppose stabilized symbolic identity. Frequency statistics, entropy measures, pattern recognition, and even information-theoretic definitions of secrecy rely on the ability to recognize repeated symbolic units and to treat them as instances of the same type. When encoding is non-canonical and symbolic identity does not stabilize, these tools do not merely lose effectiveness; their domain of applicability collapses. In such cases, the analytical framework itself no longer applies.

This perspective helps to distinguish between different notions of secrecy. In classical cryptographic models, secrecy is achieved within an established symbolic framework: symbols exist, but their relationships are obscured by randomness or computational hardness. In contrast, the boundary case analyzed here corresponds to a situation in which the symbolic framework does not arise for the observer. The resulting opacity is not due to insufficient data, noise, or complexity, but to the absence of the symbolic units required for interpretation. This condition may be described as non-recognizability rather than secrecy in the traditional sense.

Importantly, the analysis does not propose a new encryption scheme or challenge the practical value of existing cryptographic methods. Instead, it delineates a conceptual limit on what cryptographic and statistical analysis can presuppose. Within this limit, questions of optimal algorithms, key sizes, or computational resources are secondary to the more fundamental issue of whether a message exists as a message for the observer at all.

This distinction has broader implications beyond cryptography. It clarifies why certain decoding tasks—whether historical, contemporary, or hypothetical—cannot be reduced to problems of pattern extraction or learning. It also highlights the epistemic role of alphabet formation as a prerequisite for any symbolic or informational analysis. Recognizing this role allows for a clearer understanding of the conditions under which cryptographic reasoning is meaningful, and of the boundaries beyond which it ceases to apply. This observation motivates further investigation of alphabet formation as a precondition for symbolic modeling across diverse domains.

In conclusion, cryptography operates within a domain defined by the prior emergence of symbolic identity. When this emergence is prevented, the resulting opacity marks not a failure of cryptographic technique but a fundamental epistemic boundary. Identifying and articulating this boundary contributes to a more precise understanding of the scope and limits of cryptographic and information-theoretic analysis.

References

Kahn, David (1967). *The Codebreakers*. Macmillan.

Kerckhoffs, Auguste (1883). “La cryptographie militaire”. In: *Journal des sciences militaires*.

Shannon, Claude E. (1949). “Communication Theory of Secrecy Systems”. In: *Bell System Technical Journal* 28.4, pp. 656–715.