Cognitive Fingerprints: A Mathematical Framework for Statistical Profiling of Text

Adriashkin Roman
ORCID: 0009-0009-6337-1806

April 22, 2026

Abstract

We introduce a mathematical framework for representing text samples by compact finite-dimensional profiles, termed *cognitive fingerprints*. The framework is based on a feature map from token sequences into a Euclidean space of empirical statistics. This representation supports quantitative comparison of text samples through profile distances and similarity measures, as well as perturbation-based questions about stability under bounded edits. The goal of the framework is not to provide definitive judgments about a text, but to establish a reproducible mathematical basis for profile construction, profile comparison, and empirical study of structure in collections of text samples. We formulate the basic setting, define the profile representation, and record conservative stability statements under explicit modelling assumptions.

1 Introduction

Quantitative study of text often requires a representation that is compact enough for analysis, yet expressive enough to preserve regularities of interest. A common mathematical strategy is to map structured objects into finite-dimensional feature spaces and then reason about geometry, distances, and perturbations in that space. In the case of text, such a representation provides a principled way to study regularity, variation, and similarity across samples without reducing the problem to a single scalar summary. The historical background for this viewpoint touches both information-theoretic thinking and quantitative style analysis [6, 3].

The CogniPrint programme adopts this viewpoint. Its central object is the *cognitive fingerprint* of a text sample: a finite-dimensional vector of empirical statistics extracted from the text. The purpose of this object is not to act as a final judgment mechanism. Rather, it serves as a mathematical profile that enables reproducible comparison of text samples and collections of text samples. Early work on quantitative composition curves and later computational approaches to style analysis provide useful historical context for this direction [5, 1, 2].

This framing leads naturally to four questions. First, what class of feature maps yields a useful profile representation? Second, how should one compare profiles in a way that respects both magnitude and geometric orientation in feature space? Third, under what assumptions does the profile remain stable when a text sample is perturbed by a bounded number of edits? Fourth, how do aggregated profiles behave for collections of samples?

The present manuscript takes a deliberately conservative position. It does not claim universal invariance properties for arbitrary feature maps, nor does it claim that any profile representation can support definitive conclusions about authorship or source. Instead, it introduces a formal setting, states explicit assumptions, and derives basic bounds that clarify what can be proved

once a suitable feature map has been fixed. The choice to work with explicit feature maps and quantitative comparison is also compatible with general retrieval and vector-space perspectives on text representation [4].

1.1 Contributions of the manuscript

This manuscript makes the following limited but precise contributions:

1. It defines a general feature-map framework for constructing finite-dimensional text profiles.
2. It formalises the notion of a cognitive fingerprint and its normalised counterpart.
3. It introduces profile comparison through Euclidean distance and cosine similarity.
4. It states conservative stability results under explicit coordinate-wise Lipschitz assumptions.
5. It defines aggregated profiles for corpora and records the corresponding mean-profile perturbation bound.
6. It records a reproducible empirical protocol for future study without claiming completed experimental results.

1.2 Scope and limitations

The analysis in this manuscript is intentionally narrow in scope. The results are conditional on explicit assumptions about the chosen feature coordinates. They are not intended to be read as universal theorems about all conceivable text statistics. Likewise, this manuscript does not yet include a completed empirical results section. Its primary purpose is to establish a clean mathematical foundation and a disciplined vocabulary for future formal and empirical work.

2 Notation and standing assumptions

We use the following notation throughout the manuscript.

- $\mathcal{V}$ denotes a finite vocabulary.
- $\mathcal{T} = \bigcup_{n \geq 1} \mathcal{V}^n$ denotes the set of all finite token sequences over $\mathcal{V}$.
- A point $T \in \mathcal{T}$ is called a *text sample*.
- The token length of T is denoted by $|T|$.
- The Euclidean norm on $\mathbb{R}^d$ is denoted by $\| \cdot \|_2$.
- The angle-based similarity between non-zero vectors is measured by cosine similarity.
- $d_{\text{edit}}(T_1, T_2)$ denotes an edit distance on token sequences.

Two assumptions recur in the analysis.

(A1) **Coordinate stability:** each chosen feature coordinate is Lipschitz with respect to the edit metric in the regime of interest.

(A2) **Non-degeneracy:** the profile norm stays bounded away from zero when normalised comparisons are considered.

These are not universal claims about arbitrary text statistics. They are modelling assumptions about the selected feature family and the specific analysis regime.

3 Formal setting

Let $\mathcal{V}$ be a finite vocabulary, and let

$$\mathcal{T} = \bigcup_{n \geq 1} \mathcal{V}^n$$

be the set of all finite token sequences over $\mathcal{V}$.

3.1 Feature map

Let

$$\phi : \mathcal{T} \to \mathbb{R}^d$$

be a feature map. The coordinates of ϕ are measurable empirical statistics extracted from a text sample. Depending on the application, such coordinates may include bounded counts, length-normalised summaries, variation measures, or dispersion measures.

The theory developed here does not require a specific feature family, but it does require that the chosen coordinates admit explicit control under perturbation in the regime of interest.

3.2 Cognitive fingerprint

Definition 3.1 (Cognitive fingerprint)**.** *The vector*

$$\phi(T) = (\phi_1(T), \ldots, \phi_d(T))$$

is called the cognitive fingerprint *of the text sample T.*

When profile comparison should be invariant under global rescaling, one may also work with the normalised profile

$$\widehat{\phi}(T) = \frac{\phi(T)}{\|\phi(T)\|_2},$$

whenever $\phi(T) \neq 0$.

3.3 Profile comparison

For two text samples $T_1, T_2 \in \mathcal{T}$, define the Euclidean profile distance

$$D_2(T_1, T_2) = \|\phi(T_1) - \phi(T_2)\|_2.$$

If both profiles are non-zero, define the cosine profile similarity

$$S_{\cos}(T_1, T_2) = \frac{\langle \phi(T_1), \phi(T_2) \rangle}{\|\phi(T_1)\|_2 \, \|\phi(T_2)\|_2}.$$

The first quantity measures absolute separation in feature space, while the second measures angular alignment.

3.4 Perturbation model

Let $d_{\mathrm{edit}}(T_1, T_2)$ denote an edit distance on token sequences. The central perturbation question is how much the profile can change when the text sample is modified by a bounded number of edits.

Assumption 3.2 (Coordinate stability)**.** *For each coordinate ϕ_i there exists a constant $L_i \geq 0$ such that*

$$|\phi_i(T_1) - \phi_i(T_2)| \leq L_i \, d_{\mathrm{edit}}(T_1, T_2)$$

for all text pairs in the analysis regime.

This is a modelling assumption about the selected feature family. It is not asserted as a universal fact for arbitrary text statistics.

Proposition 3.3 (Basic stability statement)**.** *Under the coordinate stability assumption one obtains the bound*

$$\|\phi(T_1) - \phi(T_2)\|_2 \leq \Big(\sum_{i=1}^{d} L_i^2 \Big)^{1/2} d_{\mathrm{edit}}(T_1, T_2).$$

Proof. By coordinate stability,

$$(\phi_i(T_1) - \phi_i(T_2))^2 \leq L_i^2 d_{\mathrm{edit}}(T_1, T_2)^2.$$

Summing over $i = 1, \ldots, d$ gives

$$\|\phi(T_1) - \phi(T_2)\|_2^2 \leq \sum_{i=1}^{d} L_i^2 \, d_{\mathrm{edit}}(T_1, T_2)^2.$$

Taking square roots yields the result. □

3.5 Aggregated profiles

For a finite collection of text samples

$$\mathcal{C} = \{T^{(1)}, \ldots, T^{(N)}\},$$

one may define the empirical mean profile

$$\overline{\phi}(\mathcal{C}) = \frac{1}{N} \sum_{j=1}^{N} \phi(T^{(j)}).$$

This aggregated object is useful when the object of study is a corpus rather than a single sample.

4 Stability of normalised and aggregated profiles

The previous section gives a perturbation bound for raw profile vectors. In many applications, however, comparisons are made after normalisation, and the unit of analysis may be a collection of samples rather than a single text. We therefore record the corresponding conservative extensions.

4.1 Non-degeneracy condition

To reason about normalised profiles, one must exclude the degenerate case where the profile norm becomes arbitrarily small.

Assumption 4.1 (Non-degeneracy)**.** *There exists a constant $m > 0$ such that*

$$\|\phi(T)\|_2 \geq m$$

for all text samples considered in the analysis regime.

This assumption is not automatic. It must be checked or enforced by the particular feature design and data regime.

Proposition 4.2 (Normalised profile stability)**.** *Under the coordinate stability and non-degeneracy assumptions, there exists a constant $C > 0$ such that*

$$\|\widehat{\phi}(T_1) - \widehat{\phi}(T_2)\|_2 \leq C\, d_{\text{edit}}(T_1, T_2)$$

for all text samples in the analysis regime.

Proof sketch. The unnormalised profile map is Lipschitz by the previous proposition. The normalisation map $x \mapsto x/\|x\|_2$ is Lipschitz on the set $\{x \in \mathbb{R}^d : \|x\|_2 \geq m\}$. Composition of these two Lipschitz maps gives the result. □

Proposition 4.3 (Mean-profile perturbation bound)**.** *Let*

$$\mathcal{C}_1 = \{T_1^{(1)}, \ldots, T_1^{(N)}\}, \qquad \mathcal{C}_2 = \{T_2^{(1)}, \ldots, T_2^{(N)}\}$$

be two aligned collections of equal size. Under the coordinate stability assumption,

$$\|\overline{\phi}(\mathcal{C}_1) - \overline{\phi}(\mathcal{C}_2)\|_2 \leq \frac{1}{N} \sum_{j=1}^{N} K\, d_{\text{edit}}\big(T_1^{(j)}, T_2^{(j)}\big),$$

where

$$K = \Big(\sum_{i=1}^{d} L_i^2\Big)^{1/2}.$$

Proof. By definition of the empirical mean profile,

$$\overline{\phi}(\mathcal{C}_1) - \overline{\phi}(\mathcal{C}_2) = \frac{1}{N} \sum_{j=1}^{N} \big(\phi(T_1^{(j)}) - \phi(T_2^{(j)})\big).$$

Applying the triangle inequality and then the basic stability proposition termwise gives the result. □

4.2 Discussion of assumptions

The previous propositions should be interpreted carefully.

- The inequalities themselves are theorem-level statements once the assumptions are fixed.
- The assumptions are modelling statements about the chosen feature map and the analysis regime.
- Therefore the framework is rigorous, but only conditionally so: its strength depends on how well the selected coordinates satisfy the stated stability requirements.

This separation between formal implications and feature-design assumptions is essential for keeping the theory honest.

5 Limits of the framework

The CogniPrint framework is intended as a mathematical basis for profile construction and comparison. It is not intended as a universal inference engine.

5.1 Dependence on feature choice

Every substantive conclusion depends on the choice of feature coordinates. Different feature families can induce different geometries, different notions of closeness, and different perturbation behaviour. No claim in this manuscript should be read as feature-independent unless this is explicitly proved.

5.2 Dependence on regime

The perturbation bounds are meaningful only in the regime in which the coordinate stability assumptions hold. If a chosen statistic reacts non-smoothly to small text edits, the corresponding Lipschitz constant may be large or the assumption may fail altogether.

5.3 No definitive inference claim

The framework does not justify definitive judgments about authorship, source, or any other categorical conclusion. Its role is to define a mathematically controlled profile space and to enable reproducible comparison inside that space.

5.4 Empirical work still required

A full research programme must still answer empirical questions, including the identification of feature coordinates with meaningful stability properties, the behaviour of profile geometry under heterogeneity of length and style, the effect of normalisation procedures, and the behaviour of aggregated profiles under sampling variation.

6 Empirical protocol

This section records a conservative empirical protocol for future experiments. It is written as a reproducible plan, not as a statement of completed results.

6.1 Corpus construction principles

The empirical study should begin with explicitly defined corpora rather than ad hoc samples. At minimum, each corpus description should specify:

- the source and collection procedure;
- language and tokenisation assumptions;
- admissible length range for samples;
- inclusion and exclusion rules;
- whether the corpus is analysed at the single-sample or aggregated level.

The protocol should avoid mixing heterogeneous sources without documentation. If corpora of different genres or length distributions are compared, that heterogeneity must be stated explicitly.

6.2 Preprocessing rules

All preprocessing steps must be deterministic and documented. At minimum, the study should record:

- tokenisation procedure;
- case normalisation policy;
- punctuation handling;
- treatment of numerals, URLs, or markup;
- truncation or minimum-length rules.

No empirical comparison should be reported without freezing these preprocessing choices in advance.

6.3 Feature extraction protocol

Given a fixed preprocessing pipeline, the study should define the feature family used to build the map ϕ. The protocol should state:

- the full coordinate list;
- any normalisation applied to raw statistics;
- the admissible range of each coordinate when relevant;
- whether the resulting profile is used in raw or normalised form.

If feature ablations are performed, each ablation should be treated as a separate experimental setting.

6.4 Profile comparison metrics

The empirical protocol should specify the comparison quantities before any result is reported. The minimum set may include:

- Euclidean profile distance;
- cosine profile similarity;
- within-corpus and between-corpus dispersion summaries;
- distance between empirical mean profiles when aggregated analysis is used.

If additional comparison metrics are introduced, they should be justified mathematically or empirically rather than chosen post hoc.

6.5 Perturbation experiments

The perturbation analysis should be designed to test how profile geometry changes under controlled edits. Representative perturbation families may include bounded token insertions, bounded token deletions, bounded substitutions, and sentence-level reorderings when relevant.

For each perturbation family, the protocol should record:

- the perturbation budget;
- whether edits are random or structured;
- the number of repeated trials;
- the summary statistic used to report profile change.

The purpose of this section is not to demonstrate universal robustness, but to identify the regime in which the chosen feature family behaves in a stable and interpretable way.

6.6 Robustness analysis

The robustness layer should include at least the following checks:

- sensitivity to text length;
- sensitivity to preprocessing choices;
- sensitivity to feature ablation;
- sensitivity to corpus heterogeneity.

Whenever possible, robustness results should be reported with uncertainty summaries rather than isolated point values.

6.7 Reproducibility rule

Every empirical run should be reproducible from a documented configuration consisting of:

- corpus definition;
- preprocessing version;
- feature definition;
- comparison metric set;
- perturbation protocol;
- random seed, if randomness is present.

7 Conclusion and open problems

This manuscript establishes a conservative mathematical basis for the CogniPrint framework. Its main contribution is structural: it defines a finite-dimensional profile representation for text samples, formalises profile comparison, and records conditional stability statements under explicit assumptions.

The framework should be understood as a basis for analysis, not as a final inference mechanism. Its value lies in the fact that profile construction, profile comparison, and perturbation questions can be studied within a controlled mathematical language.

Several open problems remain central for future work:

1. identifying feature families with strong empirical stability properties;
2. deriving concentration results for empirical mean profiles under realistic sampling assumptions;
3. understanding how feature-space geometry changes across heterogeneous corpora;
4. formalising admissible normalisation procedures for robust profile comparison;
5. linking empirical protocol design more tightly to theorem-level guarantees.

A mature version of the preprint should therefore combine two elements: a mathematically explicit representation framework and a carefully controlled empirical programme. Only the combination of these two parts can justify stronger conclusions in future iterations.

A Appendix scaffold

This appendix is intentionally minimal at the current stage. In the next manuscript pass it may be used for:

- extended proof details;
- notation tables;
- admissible feature-family examples;
- reproducibility checklists;
- supplementary protocol details.

References

[1] J. F. Burrows. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Clarendon Press, Oxford, 1987.

[2] Maciej Eder, Jan Rybicki, and Mike Kestemont. Stylometry with R: A package for computational text analysis. *The R Journal*, 8(1):107–121, 2016.

[3] David I. Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 1998.

[4] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Sch"utze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008.

[5] T. C. Mendenhall. The characteristic curves of composition. *Science*, ns-9(214S):237–246, 1887.

[6] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.