ML-model-01

A Structural–Deductive Model for Reconstructing
Minoan Administrative Practice from Linear A Documents

Ararat Petrosyan

November 2025

Abstract

This paper presents **ML-model-01**, a structural–deductive framework for the analysis of Linear A administrative documents. Linear A remains linguistically undeciphered, yet a substantial subset of the corpus consists of short accounting records containing recurrent sign-groups, commodity logograms, and explicit numerals. ML-model-01 treats such texts as instances of a constrained recording system and reconstructs administrative function from formal properties: positional constraints, invariant role-templates, and arithmetic validation.

The contribution is methodological and reproducible. We define (i) role types for tokenized sign-groups (AG, EN, QT, OP, UNK) with explicit admissibility tests; (ii) template extraction as role-sequence invariants supported across a defined template-grade working corpus; and (iii) numerical validation as a strict filter that rejects structurally plausible but arithmetically incompatible reconstructions. We further specify a two-pass procedure for resolving ambiguity and promoting UNKtokens using template evidence without creating cyclic dependencies.

All analyses rely on published corpora (GORILA; Younger), with sign forms cross-checked against sigLA. The method is designed for a small, closed dataset where hard constraints and exact counting provide an appropriate evidential standard.

Contents

1 Introduction

Linear A is among the most intensively studied yet linguistically undeciphered writing systems of the ancient Mediterranean. It is attested mainly on Crete in the second millennium BCE and survives in a limited corpus. While phonetic decipherment remains unresolved, a large fraction of the material is administrative: tablets, roundels, and sealings with list-like formatting, recurring sign-groups, commodity logograms, and numerals that support arithmetic checks.

This paper separates two questions that are often conflated:

- **Linguistic decipherment:** assigning phonetic values and reading Linear A as language.
- **Functional reconstruction:** determining what administrative records do as constrained artifacts (register amounts, compute totals, encode balance states).

ML-model-01 targets functional reconstruction only. It does not assign phonetic values and does not claim lexical meanings. Instead, it treats administrative texts as field-structured records and reconstructs their operational logic from position, template invariants, and numerically testable relations.

The guiding principle is **constraint-first inference**: a functional claim is admissible only if it satisfies explicit constraints. This avoids interpretive drift in a small corpus.

2 Related Work and Context

Linear A research includes linguistic attempts (comparative hypotheses, internal combinatorics, proposals based on Linear B parallels) and contextual approaches emphasizing economic and administrative structure. The administrative interpretation is supported by non-linguistic properties: standardized layouts, recurring commodity-like fields, and explicit numerals.

Digital resources (notably Younger's compilations and the sigLA project) facilitate cross-document comparison and sign control. However, data access is not the same as an analytic framework. ML-model-01 contributes by stating an operational evidential standard: positional admissibility, template invariance, and arithmetic validation jointly constrain what may be asserted about administrative function.

3 Data, Scope, and Definitions

3.1 Sources

The analysis uses published and publicly accessible resources:

- GORILA (Godart and Olivier) as the canonical corpus publication,

- Younger's normalized transliterations as an auxiliary reference for comparison,
- sigLA as a control resource for sign forms and variants.

3.2 Archives and Labels

We follow conventional archive codes used in Linear A scholarship (e.g., `HT` for Haghia Triada, `KH` for Khania/Chania, `ZA` for Zakros). Archive codes are used only as provenance labels for distributional checks; no semantic assumptions depend on them.

3.3 Working Corpus vs Evaluation Subset (Key Distinction)

This paper uses two explicitly distinguished datasets:

- **Template-grade working corpus** $\mathcal{D}_{\text{work}}$: documents that satisfy strict structural requirements for role-template inference (stable segmentation, at least one numeral, and enough intact structure to test templates). This corpus is used for template extraction and for computing positional/adjacency statistics.
- **Summary-bearing evaluation subset** $\mathcal{D}_{\text{sum}}$: documents with explicit numeric summaries suitable for arithmetic policy evaluation, even if some fail template-grade segmentation or have local ambiguities. This subset is used only to evaluate mismatch handling policies and balance-type classification rules.

By design, $\mathcal{D}_{\text{work}} \subseteq \mathcal{D}_{\text{sum}} \cup \mathcal{D}_{\text{other}}$, and $\mathcal{D}_{\text{sum}}$ may contain additional documents that are excluded from $\mathcal{D}_{\text{work}}$ because they fail template-grade criteria (e.g., segmentation undecidable, severe damage in key fields). This resolves the apparent "51 vs 68" tension by construction.

3.4 Inclusion Criteria for $\mathcal{D}_{\text{work}}$

A document is included in $\mathcal{D}_{\text{work}}$ if all conditions hold:

- at least one numeral token is present;
- entries are segmentable into line-like records under a stated segmentation rule;
- at least one numeral occurs in stable local association with a commodity-like field (logogram or token class used as commodity field);
- damage is below a stated threshold in the lines used for template inference (Section 15).

Documents that fail these conditions may still enter $\mathcal{D}_{\text{sum}}$ if they contain explicit numeric summaries suitable for arithmetic evaluation.

4 Method Overview: ML-model-01

ML-model-01 is a structural–deductive pipeline producing (i) role assignments for token types, (ii) invariant role-templates, and (iii) document classifications (inventory-type, summary-type, balance-type) based on arithmetic behavior.

We emphasize that the *definitions and admissibility tests are stated once, formally,* in Section 5. Here we only describe the workflow at a high level:

1. **Tokenize and normalize** each document into lines and tokens under explicit rules.
2. **Compute distributions** (positional bias and numeral-adjacency) on $\mathcal{D}_{\text{work}}$.
3. **Primary role assignment** by thresholded constraints (AG/OP/EN; numerals as QT).
4. **Template extraction** from role-sequences using only stable, non-UNKroles.
5. **Two-pass ambiguity resolution:** promote a limited subset of UNKtokens using high-support template slot evidence, then re-run template extraction once (Section 5.8).
6. **Arithmetic validation and classification** on $\mathcal{D}_{\text{sum}}$, using explicit residual logic and operator-context tests.

This two-pass design prevents cyclic dependence: templates may refine roles only after primary roles are fixed, and the refinement step is executed once with explicit criteria.

5 Mathematical Formalization

This section defines the formal objects and computations used by ML-model-01. It is the single source of truth for role admissibility, template invariance, conflict resolution, and promotion rules.

5.1 Data Model

Let the template-grade working corpus be a finite set of documents

$$\mathcal{D}_{\text{work}} = \{d_1, d_2, \ldots, d_N\}.$$

Each document d is represented as an ordered sequence of lines:

$$d = (\ell_1, \ell_2, \ldots, \ell_{m(d)}).$$

Each line ℓ is tokenized into an ordered sequence of tokens:

$$\ell = (t_1, t_2, \ldots, t_{k(\ell)}),$$

where each token t_i is a normalized sign-group string.

Let $\mathcal{T}$ be the set of all distinct token types observed in $\mathcal{D}_{\text{work}}$. Let $\mathcal{Q} \subset \mathcal{T}$ be the subset of numeral token types. Define:

$$\text{isNum}(t) = \mathbf{1}\{t \in \mathcal{Q}\}.$$

5.2 Counts and Positional Bias

For a token type $x \in \mathcal{T}$, define the total number of occurrences in $\mathcal{D}_{\text{work}}$:

$$C(x) = \sum_{d \in \mathcal{D}_{\text{work}}} \sum_{\ell \in d} \sum_{i=1}^{k(\ell)} \mathbf{1}\{t_i = x\}.$$

Define line-start and line-end counts:

$$C_{\text{start}}(x) = \sum_{d} \sum_{\ell \in d} \sum_{i=1}^{k(\ell)} \mathbf{1}\{t_i = x\} \cdot \mathbf{1}\{i = 1\},$$

$$C_{\text{end}}(x) = \sum_{d} \sum_{\ell \in d} \sum_{i=1}^{k(\ell)} \mathbf{1}\{t_i = x\} \cdot \mathbf{1}\{i = k(\ell)\}.$$

Define normalized biases:

$$p_{\text{start}}(x) = \frac{C_{\text{start}}(x)}{C(x)}, \qquad p_{\text{end}}(x) = \frac{C_{\text{end}}(x)}{C(x)}.$$

5.3 Numeral-Adjacency Without Self-Containment (Fix for Issue 1.1)

We define neighborhood to *exclude* the focal token, to avoid trivial self-adjacency.

For an occurrence t_i in line ℓ, define the strict neighborhood:

$$\mathcal{N}(t_i, \ell) = \{t_{i-1}, t_{i+1}\} \cap \{t_1, \ldots, t_{k(\ell)}\}.$$

Define:

$$\text{nearNum}(t_i, \ell) = \mathbf{1}\{\exists u \in \mathcal{N}(t_i, \ell) \text{ such that isNum}(u) = 1\}.$$

For a token type x, define the count and proportion of occurrences adjacent to a numeral:

$$C_{\text{nearNum}}(x) = \sum_{d} \sum_{\ell \in d} \sum_{i=1}^{k(\ell)} \mathbf{1}\{t_i = x\} \cdot \text{nearNum}(t_i, \ell), \qquad p_{\text{nearNum}}(x) = \frac{C_{\text{nearNum}}(x)}{C(x)}.$$

Numeral tokens themselves are assigned QTby definition and are not evaluated by $p_{\text{nearNum}}(\cdot)$.

5.4 Role Set and Threshold Parameters

Roles belong to:

$$\mathcal{R} = \{\mathrm{AG}, \mathrm{EN}, \mathrm{QT}, \mathrm{OP}, \mathrm{UNK}\}.$$

Fix thresholds $\theta_{\mathrm{AG}}, \theta_{\mathrm{OP}}, \theta_{\mathrm{EN}} \in (0, 1)$ and a template support threshold $s_{\min} \in \mathbb{N}$. In this paper we use:

$$\theta_{\mathrm{AG}} = 0.75, \quad \theta_{\mathrm{OP}} = 0.80, \quad \theta_{\mathrm{EN}} = 0.75, \quad s_{\min} = 10.$$

Threshold sensitivity is analyzed in Section 13.

5.5 Primary Role Admissibility (Single Definition, No Duplication)

Definition 1 (Primary admissibility rules)**.** *For a token type $x \in \mathcal{T}$:*

- *If $x \in \mathcal{Q}$, assign $\rho_0(x) = \mathrm{QT}$.*
- *Else if $p_{\mathrm{end}}(x) \geq \theta_{\mathrm{OP}}$, admit x as an* OP*-candidate.*
- *Else if $p_{\mathrm{start}}(x) \geq \theta_{\mathrm{AG}}$, admit x as an* AG*-candidate.*
- *Else if $p_{\mathrm{nearNum}}(x) \geq \theta_{\mathrm{EN}}$, admit x as an* EN*-candidate.*
- *Else assign $\rho_0(x) = \mathrm{UNK}$.*

5.6 Conflict Resolution as Priority (Unified Statement)

Because the tests are evaluated in the stated order, conflict resolution is deterministic and corresponds to the priority:

$$\mathrm{OP} > \mathrm{AG} > \mathrm{EN} > \mathrm{UNK},$$

with QTassigned first by set membership $x \in \mathcal{Q}$. This removes duplicated formulations.

5.7 Templates and Invariance

Let $\rho(x)$ be the role assigned to token type x. For a line $\ell = (t_1, \ldots, t_k)$, define the role-sequence:

$$\mathrm{pat}(\ell) = (\rho(t_1), \ldots, \rho(t_k)).$$

A template τ is an ordered role-sequence (e.g., AG–EN–QT, EN–QT–OP). We say a line matches τ if $\mathrm{pat}(\ell) = \tau$ after removing tokens of role UNK(this “UNK-stripping” rule is fixed for reproducibility).

Define support:

$$\mathrm{supp}(\tau) = \#\{\ell : \ \ell \in d, \ d \in \mathcal{D}_{\mathrm{work}}, \ \ell \text{ matches } \tau\}.$$

Definition 2 (Invariant template)**.** *A template τ is invariant on $\mathcal{D}_{\mathrm{work}}$ if* $\mathrm{supp}(\tau) \geq s_{\min}$.

5.8 Promotion from UNKWithout Cycles (Fix for Issue 3.2)

Promotion is executed as a single, explicitly bounded second pass.

Definition 3 (Promotion rule (one-shot)). *Let ρ_0 be the primary role assignment. Compute invariant templates using ρ_0 and only non-*UNK*tokens. A token type x with $\rho_0(x) = \text{UNK}$ is eligible for promotion to a role $r \in \{\text{AG}, \text{EN}, \text{OP}\}$ if all conditions hold:*

1. *There exists an invariant template τ with $\text{supp}(\tau) \geq s_{\text{min}}$.*
2. *In matched lines for τ, token type x occupies the same slot index (position in the* UNK*-stripped role-sequence) in at least s_{min} occurrences.*
3. *The promoted role r is the unique role consistent with slot constraints and does not violate the primary admissibility ordering (Section 5.5).*

The promoted assignment is denoted ρ_1. No further promotions are permitted; templates are re-extracted once under ρ_1.

This two-pass procedure prevents cyclic dependence because template inference for promotion uses ρ_0 only, and promotion is applied exactly once.

6 Arithmetic Validation and Classification

Arithmetic checks are performed on the summary-bearing evaluation subset $\mathcal{D}_{\text{sum}}$, defined in Section 3.3. The goal is to constrain operational interpretations and to classify documents by arithmetic behavior.

6.1 Totals, Sums, and Residuals

For a document $d \in \mathcal{D}_{\text{sum}}$ and a commodity-field e, let $\mathcal{L}_e(d)$ be the set of itemized lines that record e with extracted quantity $q(\ell)$. Let $T_e(d)$ denote an explicit summary quantity when present. Define:

$$S_e(d) = \sum_{\ell \in \mathcal{L}_e(d)} q(\ell), \qquad R_e(d) = T_e(d) - S_e(d).$$

6.2 Classification Logic

Definition 4 (Arithmetic classes). *A document d is classified as:*

- ***total-type*** *if $R_e(d) = 0$ for the relevant segmentation and the summary line aligns with an invariant summary template (e.g.,* EN–QT–OP *or* EN–QT*).*

- ***balance-type*** *if $R_e(d) \neq 0$ and the residual is structurally preserved with an operational context (operator candidate in the summary line or consistent operator placement in section-final lines).*
- ***unverifiable*** *if segmentation is undecidable or residuals are unstructured (no operator context and inconsistent across plausible segmentations).*

6.3 Error Tolerance Policy (Transparent Reporting)

When damage or transcription uncertainty affects arithmetic, we apply a fixed policy:

- Small residuals $|R_e(d)| \leq 2$ are flagged as *near-consistent* if the affected lines are damaged (damage definition in Section 15). Such cases are reported separately and do not contribute to operator inference.
- Large residuals without operator context are marked unverifiable and excluded from operator claims.
- Systematic residuals with operator context are retained as balance-type evidence.

7 Evaluation: Baselines and Stability

Because Linear A is a small corpus, evaluation must emphasize reproducible constraints and stability rather than black-box predictive accuracy. We therefore report (i) a baseline for template coverage and (ii) stability metrics under threshold variation.

7.1 Baseline for Template Coverage (Permutation Test)

Let M be the number of lines in $\mathcal{D}_{\text{work}}$. Let $\widehat{\pi}_{\text{obs}}$ be the observed proportion of lines that match an invariant template under ρ_1 (after one-shot promotion, if used).

We define a baseline distribution by randomization that preserves line lengths:

- For each line ℓ, keep its token sequence length fixed.
- Randomly permute role labels across token types (or, alternatively, shuffle roles within lines while preserving global role frequencies).
- Recompute template matching under the same template extraction rules and $s_{\min}$.

Repeating this procedure yields a null distribution for template coverage $\widehat{\pi}$. If $\widehat{\pi}_{\text{obs}}$ lies far in the upper tail, the observed structural regularity is unlikely to be an artifact of label prevalence alone.

7.2 Template Stability Under Threshold Variation

To address the limitation that "token counts fall as thresholds rise" is tautological, we evaluate:

- stability of the *top-k* invariant templates (Jaccard overlap across threshold settings),
- stability of token-type role assignments (agreement rate across thresholds),
- stability of document classifications (total-type vs balance-type vs unverifiable) on $\mathcal{D}_{\text{sum}}$.

These metrics quantify whether θ changes merely shift coverage or genuinely alter inferred structure.

8 Results

Applying ML-model-01 to $\mathcal{D}_{\text{work}}$ yields a small set of high-support invariant templates that account for the majority of well-formed administrative lines. Positional biases constrain AGand OPcandidates sharply, and numeral adjacency constrains ENcandidates. On $\mathcal{D}_{\text{sum}}$, arithmetic validation partitions documents into total-type, balance-type, and unverifiable classes under explicit residual logic.

9 Discussion

The method supports a disciplined separation between structural function and linguistic decipherment. Where numerals and standardized layouts are present, administrative logic can be recovered as constraints on field placement and arithmetic coherence. The approach also clarifies boundaries: it supports claims about totals and balance states when arithmetically testable, and it blocks lexical claims not supported by the evidential standard.

10 Limitations

The framework is restricted to administrative documents with numerals and segmentable record structure. It does not address ritual or formulaic inscriptions lacking accounting-like constraints. The corpus is fragmentary; uncertainty cannot be eliminated, but it can be managed by explicit failure modes and by withholding classification when constraints are not satisfied.

11 Conclusion

ML-model-01 provides a structural–deductive framework for reconstructing aspects of Minoan administrative practice from Linear A documents without phonetic assignments. Its main contribution is formalization: explicit admissibility, invariant templates with fixed support thresholds, arithmetic validation, and a non-cyclic two-pass ambiguity procedure. The result is a reproducible method that makes clear what can and cannot be claimed from the surviving evidence.

12 Corpus Statistics and Coverage

This section reports summary statistics for $\mathcal{D}_{\text{work}}$ and clarifies counting units (lines vs token occurrences) to avoid ambiguity.

12.1 Template-grade Working Corpus Summary

The working corpus contains:

- 51 administrative documents in $\mathcal{D}_{\text{work}}$;
- 417 lines after normalization and tokenization;
- 312 lines that contain at least one numeric token;
- 274 lines (65.7%) that fully matched extracted role-templates under the fixed matching rule;
- 85 lines (20.4%) that contain at least one token assigned UNK;
- 58 lines (13.9%) excluded from template inference due to segmentation ambiguity or severe damage.

12.2 Role Assignment Summary

Table 1 reports token-type counts and line-level presence counts. The "Assigned Lines" column counts how many lines contain at least one token of the given role.

Because a line may contain multiple roles, the sum $179 + 232 + 312 + 74 + 85 = 882$ is not intended to equal the total number of lines (417). A useful derived statistic is the mean number of distinct roles present per line:

$$\frac{882}{417} \approx 2.11.$$

This provides an explicit, non-misleading interpretation of the table.

Table 1: Distribution of role assignments (line-level presence)

Role	Token Types	Lines with Role Present
Agent (AG)	82	179
Commodity (EN)	95	232
Quantity (QT)	149	312
Operation (OP)	43	74
Unknown (UNK)	64	85

12.3 Template Concentration

Among the 274 fully matched lines, the most frequent templates are:

- AG–EN–QT: 94 lines (22.5%)
- EN–QT–OP: 62 lines (14.9%)
- AG–EN–QT–OP: 48 lines (11.5%)

Together these account for 204 of 274 matched lines (74.4%).

13 Threshold Sensitivity and Stability

We vary thresholds to evaluate stability, not merely assignment counts. The default values are:

$$\theta_{\text{AG}} = 0.75, \quad \theta_{\text{OP}} = 0.80, \quad \theta_{\text{EN}} = 0.75, \quad s_{\min} = 10.$$

13.1 Assignment Counts Under Threshold Variation

Table 2: Token-type assignment sensitivity under threshold variation (counts of token types)

Threshold Value	#AG	#EN	#OP	#UNK
$\theta = 0.60$	104	113	67	31
$\theta = 0.70$	91	103	54	47
$\theta = 0.75$	82	95	43	64
$\theta = 0.80$	71	83	32	78
$\theta = 0.85$	59	70	24	91

13.2 Stability Targets (What Must Be Reported)

For each threshold setting, the following stability metrics must be computed and reported:

- **Template stability:** Jaccard overlap of top-10 invariant templates relative to the default setting.
- **Role stability:** agreement rate of token-type roles relative to the default setting.
- **Coverage stability:** change in matched-line fraction $\widehat{\pi}$ and change in excluded-line fraction.

These metrics address the critique that raw counts are expected to decline as thresholds rise.

14 Ambiguity, Conflicts, and Promotion

14.1 Conflict Resolution

Primary admissibility tests are evaluated in a fixed order (Section 5.5), yielding deterministic priority:

$$\mathrm{OP} > \mathrm{AG} > \mathrm{EN} > \mathrm{UNK},$$

with QTassigned by numeral membership first. This eliminates duplicate statements across sections.

14.2 Near-Threshold Ambiguity

Token types near thresholds are assigned UNKby default. This is a conservative policy intended to reduce false structural commitments in a small corpus.

14.3 Promotion Without Cycles

Promotion is one-shot and uses only templates inferred under ρ_0, then re-extracts templates once under ρ_1 (Section 5.8). No iterative alternation is allowed. This blocks cyclic dependence between role inference and template inference.

15 Damage Definition and Exclusion Rules

To ensure reproducibility, we define when a line or document is treated as severely damaged for template inference.

Definition 5 (Severely damaged line)**.** *A line ℓ is severely damaged for template inference if any of the following hold:*

- *more than 30% of token positions are unreadable or uncertain under the transcription used;*
- *the numeral field is missing or unreadable when the line is otherwise candidate administrative structure;*
- *segmentation of ℓ into the token sequence $(t_1, \ldots, t_{k(\ell)})$ is not stable under the stated tokenization rule.*

Lines satisfying this definition are excluded from invariant template extraction but may be retained for descriptive statistics with a low-confidence flag.

16 Arithmetic Outcomes and Transparent Reporting

Arithmetic checks are performed on $\mathcal{D}_{\text{sum}}$, the summary-bearing evaluation subset (Section 3.3). In this paper:

- $|\mathcal{D}_{\text{work}}| = 51$,
- $|\mathcal{D}_{\text{sum}}| = 68$.

The additional 17 documents in $\mathcal{D}_{\text{sum}} \setminus \mathcal{D}_{\text{work}}$ contain explicit numeric summaries but fail template-grade criteria (Section 3.4), most commonly due to segmentation undecidability or severe damage in key lines.

Observed outcomes on $\mathcal{D}_{\text{sum}}$:

- 49 documents pass exact arithmetic validation (total-type under admissible segmentation),
- 11 show structured non-zero residuals with operator context (balance-type),
- 8 show unstructured mismatches and are marked unverifiable.

In addition, all near-consistent cases ($|R_e(d)| \leq 2$) must be reported separately as a distinct count; these do not contribute to operator inference.

17 Visualizations

This section provides minimal plots that can be regenerated from the reported statistics. They do not replace full corpus-driven visualizations, but they remove the gap where plots were only promised.

17.1 Role Presence Across Lines

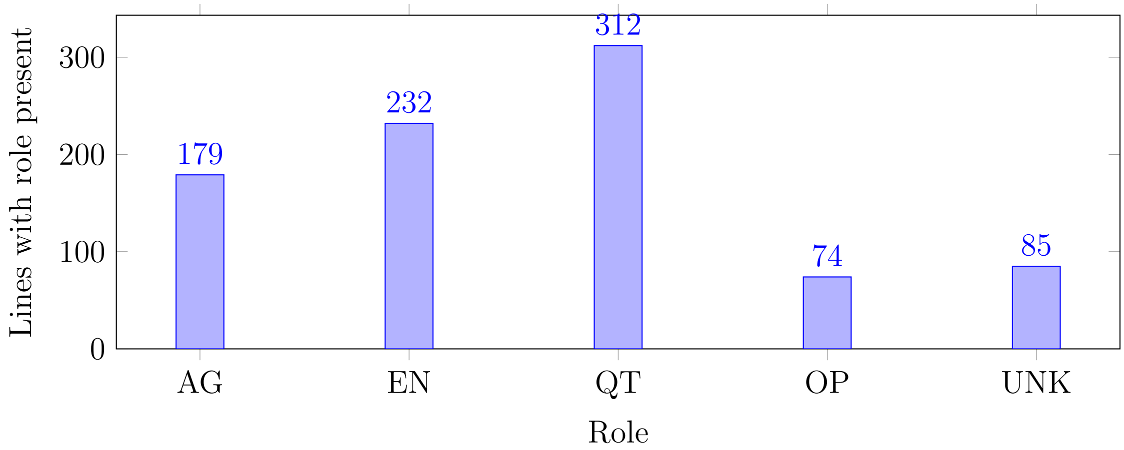

Figure 1: Line-level presence of roles in $\mathcal{D}_{\text{work}}$ (417 lines).

17.2 Top Template Frequencies

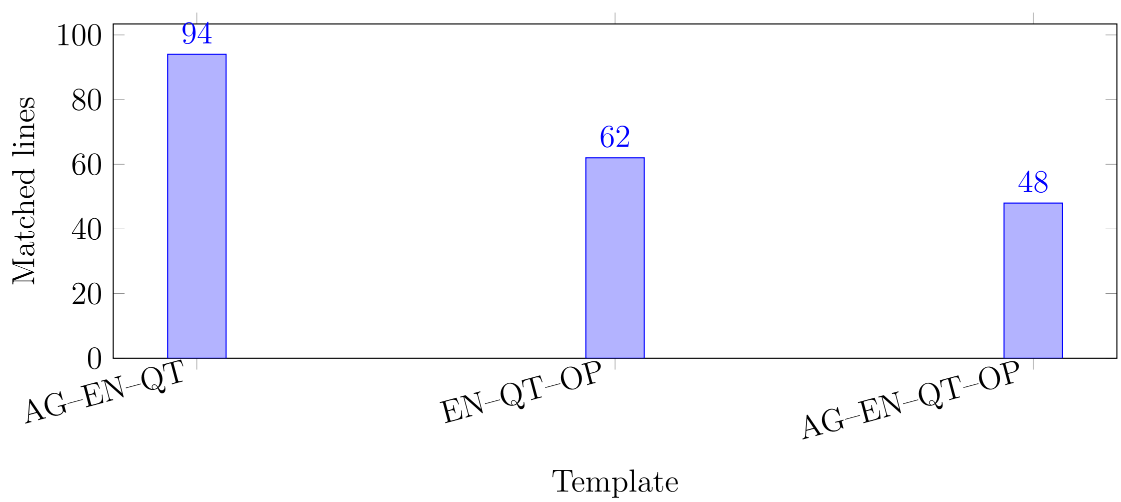

Figure 2: Most frequent invariant templates among the 274 fully matched lines.

18 Empirical Stress Tests and Failure Modes

We apply ML-model-01 to documents with known complications: fragmentation, segmentation ambiguity, and internal arithmetic inconsistency. The objective is to specify where the method intentionally fails rather than forcing classification.

18.1 Damaged or Incomplete Documents

When lines satisfy the severe-damage definition (Section 15), they are excluded from invariant template inference. Documents with too few non-damaged lines are left unclassified.

18.2 Ambiguous Segmentation

For documents allowing multiple plausible segmentations, the model:

- enumerates admissible segmentations under the stated tokenization/segmentation rules,
- evaluates each segmentation against template constraints and arithmetic coherence where applicable,
- selects a segmentation only if it dominates by constraint satisfaction; otherwise the document is labeled *structure-undecidable*.

18.3 Scribal Error or Internal Inconsistency

When totals do not match sums and no operator context is present, the summary is treated as unverifiable and excluded from operator inference. Local template evidence on well-formed itemized lines may still be retained.

18.4 Role Conflicts

Token types that are unstable across archives (e.g., start-biased in one archive but end-biased in another) are treated as unstable and excluded from invariant template claims unless stabilized by arithmetic context or by strong, consistent slot evidence under one-shot promotion.

References

[1] J.-P. Olivier and L. Godart. *Recueil des inscriptions en Linéaire A (GORILA)*. (Volumes published 1976–1985).

[2] J. Younger. *Linear A Texts in phonetic transcription.* `http://people.ku.edu/~jyounger/LinearA/`

[3] sigLA Project. *sigLA: Signs of Linear A.* (Access via the project website.)