Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Determinative Detection in Egyptian Hieroglyphic Sequences: A Sequence-Labelling Benchmark and Multi-Agent LLM Baseline

Daniel Cruz-Cavalieri

daniel.cavalieri@ifes.edu.br

Instituto Federal do Espírito Santo

Carlos Gracia-Zamacona

Universidad de Alcalá

Jorke Grotenhuis

Universidad de Alcalá

Paulo Sergio dos Santos

Instituto Federal do Espírito Santo

Sira Elena Palazuelos-Cagigas

Universidad de Alcalá

Research Article

Keywords:

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10580069/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Determinative Detection in Egyptian Hieroglyphic Sequences: A Sequence-Labelling Benchmark and Multi-Agent LLM Baseline

Daniel Cruz-Cavalieri[1*], Carlos Gracia-Zamacona[2], Jorke Grotenhuis[2], Sira E. Palazuelos-Cagigas[3], Paulo Sergio dos Santos-Junior[4]

[1*]Department of Automation and Control Engineering, Instituto Federal do Espírito Santo (IFES), Serra, Espírito Santo, 29173-087, Brazil.
[2]Departamento de Historia y Filosofía, Universidad de Alcalá, Alcalá de Henares, Madrid, 28801, Spain.
[3]Departamento de Electrónica, Universidad de Alcalá, Alcalá de Henares, Madrid, 28871, Spain.
[4]Department of Information Systems, Instituto Federal do Espírito Santo (IFES), Serra, Espírito Santo, 29173-087, Brazil.

*Corresponding author(s). E-mail(s): daniel.cavalieri@ifes.edu.br;
Contributing authors: carlos.gracia@uah.es; jorke.grotenhuis@uah.es; sira.palazuelos@uah.es; paulo.junior@ifes.edu.br;

Abstract

Determinatives, silent classifier signs that mark semantic fields and signal word boundaries, are a central feature of the ancient Egyptian writing system, and their automatic detection is a prerequisite for lemmatization, morphological parsing, and translation. Yet the task has received little dedicated attention as a language-processing problem, and no shared benchmark exists. This paper makes three contributions. First, we formulate determinative detection in sequences encoded in Manuel de Codage notation as a sequence-labelling task and release an expert-annotated benchmark of 99 hieroglyphic sequences, structured around a knowledge base of determinative signs organised into semantic categories following the Gardiner Sign List. Second, we propose a multi-agent method combining large language model (LLM) reasoning with deterministic domain constraints: two independent LLM agents analyse each sequence from semantic and structural perspectives, and a deterministic scoring module aggregates their evidence

through a five-tier sign categorisation and context-dependent thresholds. Third, we evaluate the method under a nested cross-validation protocol in which every sequence serves as a held-out test case exactly once, with inner Bayesian optimisation of the scoring parameters. The method attains a macro-averaged F1 of **81.94% ± 2.19%** and a micro-averaged F1 of **87.50% ± 2.50%**, the low cross-fold variance quantifying its robustness. Beyond a single prediction, it produces ranked alternative interpretations with explicit promotion conditions and natural-language justifications, making it a transparent, collaborative resource for computational Egyptology, a low-resourced ancient language, rather than a black-box classifier.

Keywords: determinative detection, Egyptian hieroglyphs, annotated corpus, sequence labelling, multi-agent systems, large language models

1 Introduction

The ancient Egyptian hieroglyphic writing system, in continuous use for over three millennia, is one of the most complex and enduring communication systems in human history. Unlike purely alphabetic scripts, it employs a hybrid architecture in which signs serve multiple, context-dependent functions: the same hieroglyph may operate as a phonogram (encoding sound), an ideogram (encoding word meaning), or a *determinative*, a silent, unpronounced sign appended to the end of a word to indicate its semantic field to the reader (Allen, 2000; Gardiner, 1957). These determinatives constitute a pervasive and functionally essential component of the writing system. Nearly every word in the script is followed by one or more such signs, which serve to signal word boundaries in an otherwise continuous text, assign words to semantic fields, and encode semantic uses of Egyptian words, including disambiguating homophonous words (Gracia Zamacona, 2001).

The scholarly understanding of determinatives has evolved considerably sinceChampollion (1836) first coined the term *déterminatif* to describe these signs. In 1995, Orly Goldwasser's influential work *From Icon to Metaphor* (Goldwasser, 1995) opened a new trend in the semiotic analysis of the hieroglyphic writing system by exploiting the opposition pictorial vs. phonetic, inspired by Jakobson (1956). Over the past two decades, a new interpretation claims that determinatives constitute a rule-governed classifier system (Goldwasser, 2002,0; Goldwasser and Grinevald, 2012; Goldwasser, 2023; Goldwasser and Soler, 2024; Lincke and Kammerzell, 2012), which encodes information about conceptual categories, discourse pragmatics, and grammatical relations. This approach has extended the comparative perspective to Sumerian cuneiform and ancient Chinese script systems (Goldwasser and Handel, 2024). Appealing as the determinatives-are-classifiers theory may be, it has been approached through acontextualised studies basedon dictionary data or sign extraction from texts, without taking the specific context of each token into account. Furthermore, the boundary between the "phonetic" and the "pictorial" may not be straightforward (Vernus, 2003), and the material component of writing in ancient Egypt is not considered in this theory (Fischer, 1986; Gracia Zamacona, 2024; Vernus, 2022).

The need to test the classifiers theory against sources by performing in-context analyses has been noted (Gracia Zamacona, 2001,1,1,2; McDonald, 2002a,0,0,0,1) because determinatives clearly show high sensitivity to context. A telling example is the very frequent and polysemic verb *ꜣpw* 'to guard, wait, avoid, be late, distrust', which is attested with eleven different determinatives, including absence of any determinative, in the Coffin Texts, according to the specific context of different tokens (Gracia Zamacona, 2016). The amount of work implied by this task is too large to be done manually by specialists. The recent perspective of automating the OCR of ancient Egyptian hieroglyphs (Fuentes-Jimenez et al., 2025) has opened the way to automatically spot determinatives and segment written words within the unsegmented written chain (Gracia Zamacona, 2015).

In parallel with these theoretical advances, computational approaches to ancient languages have undergone a transformation driven by advances in artificial intelligence and machine learning. Sommerschield et al. (2023), in a comprehensive survey, identified a taxonomy of tasks spanning digitisation, restoration, attribution, linguistic analysis, textual criticism, translation, and decipherment. However, computational research on ancient Egyptian remains in its infancy, with existing work concentrated almost entirely on technical tasks such as optical character recognition. Within the specific domain of Egyptian hieroglyphs, early computational work addressed individual sign recognition through image retrieval approaches (Franken and van Gemert, 2013) and, more recently, through deep learning. Barucci et al. (2021) proposed GlyphNet, a dedicated convolutional neural network for hieroglyph image classification that outperformed standard architectures such as ResNet-50, Inception-v3, and Xception. Subsequent work extended these approaches to segmentation using Mask R-CNN (Barucci et al., 2023) and hieroglyphic text recognition at scale using ensemble ConvNeXt backbones (Fuentes-Ferrer et al., 2025).

Beyond recognition and classification, recent work has addressed higher-level NLP tasks. De Cao et al. (2024) presented a Hieroglyphic Transformer for translating ancient Egyptian into German and English, built on the M2M-100 framework and trained on data from the Thesaurus Linguae Aegyptiae (TLA). Sahala and Lincke (2024) developed neural models for lemmatisation and POS-tagging of Earlier Egyptian and Demotic. These advances nevertheless, computational tools for the Egyptian writing system remain predominantly focused on visual sign recognition and downstream translation, leaving a significant gap in the analysis of the functional roles that signs play within words, precisely the task that determinative identification requires.

The identification of determinatives is a critical prerequisite for all subsequent stages of hieroglyphic text processing. Correct identification determines the boundary between the ideo-phonographic and the semantic-segmenting portions of a word, thereby affecting lemmatisation, morphological parsing, and ultimately translation. Yet this task has received remarkably little computational attention. Barucci et al. (2021) explicitly listed recognition of determinatives and their semantic field among the open problems in computational Egyptology. Sommerschield et al. (2023) found no computational works tackling classifiers or determinatives in any ancient script. The first steps toward addressing this gap were taken by Nikolaev et al. (2024), who formulated classifier identification in ancient Egyptian as a low-resource sequence-labelling

task and implemented neural models trained on Coffin Texts annotated through the iClassifier platform (Goldwasser et al., 2019; Harel et al., 2024). Their approach, framed as a binary token-level classification problem (classifier vs. non-classifier), achieved promising results despite modest training data, but operated at the word level with frequency-based and neural sequence-labelling methods, without incorporating the rich semantic and structural cues that Egyptological theory identifies as central to determinative function.

This paper addresses the determinative identification gap with a fundamentally different approach. We present a multi-agent framework for automatic determinative detection in sequences encoded in MdC notation (Buurman et al., 1988), combining LLM reasoning with deterministic domain constraints through a pipeline of specialised agents orchestrated by LangGraph. Unlike prior work that relies on pattern-matching against annotated training corpora, the proposed method leverages the contextual reasoning capabilities of LLMs to perform semantic and structural analysis of hieroglyphic sequences, grounded in a curated knowledge base of over 80 determinatives organised into nine semantic categories with a five-tier sign categorisation scheme and context-dependent validation thresholds. Two independent LLM agents, semantic and structural, analyse each sequence from complementary perspectives without exchanging intermediate results, ensuring that their agreement constitutes genuine evidential convergence. A third LLM agent synthesises their outputs, and a deterministic scoring module applies rule-based constraints and evidence triangulation across three independent sources (semantic, structural, and rule-based) to produce the final classification.

The framework introduces several methodological innovations. First, it departs from the standard single-output classification paradigm by generating ranked alternative interpretations with explicit promotion conditions, specifying the minimal evidence that would favour an alternative over the main interpretation, thereby transforming competing hypotheses into testable scholarly propositions. Second, the system produces full natural-language justifications and rejected-candidate reasoning for every classification decision. Third, the hybrid architecture, combining LLM contextual flexibility with deterministic symbolic reasoning, addresses hallucination and inconsistency risks inherent in applying language models to specialised domains.

Evaluated on a corpus of 99 expert-annotated sequences under a nested cross-validation protocol 5-fold outer cross-validation for generalisation estimation, with each outer training partition driving an inner 5 $\times$2 cross-validation that performs Bayesian optimisation of the deterministic scoring module's parameters. The system achieves a macro-averaged F1 of 81.94% $\pm$2.19% and a micro-averaged F1 of 87.50% $\pm$2.50% across the five held-out test folds. Macro-precision and macro-recall are both above 86 %. Because every sequence in the corpus is held out as a test case in exactly one fold, the standard deviations directly quantify the system's robustness to changes in the training distribution. Performance analysis reveals a complexity paradox consistent across all folds: information-rich sequences with three or more determinatives are detected most reliably, while sparse single-determinative and empty sequences remain the residual challenge.

The remainder of this paper is organised as follows. Section 2 highlights the contributions. Section 3 describes the methodology. Section 4 presents experimental results. Section 5 discusses the findings, analyses the complexity paradox, and addresses limitations. Section 6 concludes the paper.

2 Contributions

This work makes the following contributions, organised around the task, the resource, and the method:

1. **Task formulation.** We formulate the automatic detection of determinatives in Egyptian hieroglyphic sequences encoded in Manuel de Codage notation as a sequence-labelling problem. To our knowledge this task, although explicitly identified as an open problem (Barucci et al., 2021), has not previously been given a formal computational definition or been addressed by any published system.
2. **Expert-annotated benchmark.** We release a benchmark of 99 hieroglyphic sequences annotated by expert Egyptologists, structured around a curated knowledge base of over 80 determinative signs organised into nine semantic categories with a five-tier sign categorisation scheme. The benchmark, the knowledge base, and the accompanying evaluation protocol are publicly available so that the task can be directly reused and extended by other researchers working on ancient and low-resource languages.
3. **Multi-agent method.** We propose a multi-agent architecture orchestrated through LangGraph (LangChain, Inc., 2024) that combines independent semantic and structural LLM agents with a deterministic scoring module, implementing a hybrid reasoning paradigm where neural contextual flexibility is anchored by symbolic domain constraints through the five-tier categorisation system and context-dependent validation thresholds. The method includes a deterministic word boundary segmentation component that infers word boundaries and generates contiguous suffix span candidates, following the Egyptological principle that determinatives occupy word-final positions.
4. **Rigorous evaluation.** We adopt a nested cross-validation protocol in which the deterministic scoring parameters are selected by a Bayesian inner loop (5 $\times$2 CV with Tree-structured Parzen Estimator sampling) on each outer training partition, and the full multi-agent pipeline is then evaluated on each held-out outer fold. Every sequence in the corpus contributes to the final estimate as a held-out test case exactly once, and the cross-fold variance directly quantifies the robustness of the method to changes in the training distribution.
5. **Transparent, contestable outputs.** We depart from the single-output classification paradigm by generating ranked alternative interpretations with explicit promotion conditions, together with natural-language justifications and rejected-candidate reasoning. This acknowledges the interpretive multiplicity inherent in Egyptological analysis, transforms system outputs into testable scholarly hypotheses, and makes the method a collaborative tool for domain experts rather than a black-box classifier.

3 Methodology

3.1 Task Formulation

We cast determinative detection as a sequence-labelling problem over hieroglyphic sequences encoded in Manuel de Codage (MdC) notation. Given an input sequence $S = \langle s_1, s_2, \ldots, s_n \rangle$ of n Gardiner sign codes obtained by tokenising the MdC string, the task is to assign to each position s_i a binary label $y_i \in \{0, 1\}$, where $y_i = 1$ indicates that s_i functions as a determinative in its context and $y_i = 0$ indicates any other function (phonogram, ideogram, phonetic complement, or grammatical marker). The output is the set of positions labelled as determinatives, which may be empty when a sequence contains no determinative.

Three properties distinguish this task from generic sequence labelling and motivate the design choices described in the remainder of this section. First, the label of a sign is *context-dependent*: the same Gardiner code may act as a determinative in one word and as a phonogram in another, so position-independent lookup is insufficient. Second, determinatives are governed by *positional regularities*, typically occupying word-final positions, which makes word boundary information an important signal. Third, the ground truth admits *interpretive multiplicity*: for some sequences more than one labelling is defensible on Egyptological grounds, which motivates producing ranked alternative interpretations rather than a single hard decision.

Evaluation of a predicted labelling against the expert-annotated ground truth is performed at the token level, and sequences whose ground truth contains no determinative are treated as a distinct outcome category so that correct abstention is neither rewarded nor penalised as an ordinary positive prediction (Section 3.12).

3.2 Architecture Overview

The proposed framework implements a sequential multi-agent architecture orchestrated through LangGraph (LangChain, Inc., 2024), a Python library for constructing stateful, multi-actor workflows with LLMs. The architecture follows a pipeline design in which a hieroglyphic sequence encoded in MdC notation traverses five processing stages: (1) deterministic preprocessing, (2) word boundary segmentation, (3) parallel semantic and structural analysis by independent LLM agents, (4) evidence synthesis by a third LLM agent, and (5) deterministic post-processing with scoring and final selection. Figure 1 provides a comprehensive diagram of the complete pipeline. All source code, configuration files, and experimental scripts required to reproduce the results are publicly available at: https://github.com/dcruzcavalieri/Multi_Agent_Framework_Egyptian_Hieroglyphic_Sequences

The design rests on three foundational principles, informed by recent advances in multi-agent LLM systems (Chen et al., 2024; Guo et al., 2024; Hong et al., 2024; Wu et al., 2023). First, *agent independence*: the semantic and structural analysis agents operate on the same input but from complementary perspectives, without exchanging intermediate results, thereby producing uncorrelated evidence streams. Second, *hybrid reasoning*: LLM-based contextual analysis is combined with deterministic rule-based constraints, leveraging the complementary strengths of neural flexibility and symbolic

consistency. Third, *evidence triangulation*: the deterministic scoring module aggregates evidence from three independent sources: semantic; structural; and rule-based, weighting multi-source agreement to produce robust final classifications.

A shared state object (implemented as a `TypedDict`) accompanies the sequence throughout the pipeline, accumulating preprocessing artefacts, agent outputs, and scoring results. The LangGraph framework provides explicit control flow through a directed acyclic graph of processing nodes, enabling graceful degradation when individual agents fail and supporting deterministic replay for debugging and evaluation.

3.3 Input Representation

The proposed method receives as input a single Egyptian hieroglyphic sequence encoded in MdC notation (Buurman et al., 1988; Nederhof, 2013,2), the standard digital transcription format in computational Egyptology. MdC represents hieroglyphic signs as alphanumeric Gardiner codes (e.g., `A1` for the seated man, `D54` for walking legs) and encodes spatial relationships through separators: "`-`" for linear juxtaposition, "`:`" for vertical stacking, and "`*`" for tight grouping (ligature). A typical input such as `T18-S29-(D54*A40):Z2` or  ('retinue') represents a compound word containing an ideogram, a phonogram, two determinatives, and a plural marker.

3.4 Preprocessing Module

The preprocessing module applies six deterministic steps to transform the raw MdC string into a richly annotated representation suitable for agent consumption:

- *Tokenisation* splits the MdC string at separator characters, producing an ordered list of Gardiner sign codes;
- *Structure extraction* identifies hieroglyphic blocks, vertical stacks, ligatures, and final positions within each block, structural features that carry information about sign function;
- *Rule-based baseline generation* applies a curated knowledge base of over 80 known determinatives combined with position heuristics to produce an initial set of candidate determinatives;
- The *candidate information* step retrieves, for each token, any matching entry from the knowledge base, including semantic meaning, Gardiner category, usage context, and associated keywords;
- *Few-shot context* assembly selects five expert-annotated examples, provided by the Egyptologist co-authors, that illustrate diverse determinative patterns for inclusion in agent prompts;
- Finally, *initial state assembly* populates the shared `AgentState` dictionary with all preprocessing artefacts and injects it into the LangGraph workflow.

3.5 Knowledge Base

The determinative knowledge base contains over 80 entries organised into nine semantic categories, following the Gardiner Sign List (Gardiner, 1957): Persons (A1, B1, etc.), Animals (E1, E23, etc.), Actions (D54, D40, etc.), Divinity (A40, R8, etc.),

Nature (N1, N5, etc.), Buildings (O1, O49, etc.), and Abstracts (Y1, Aa2, etc.). Signs from Gardiner's F section (Parts of Mammals) are mapped to their specific semantic fields rather than to a generic animal context; for instance, F4 (forepart of lion) is associated with the semantic fields of "front," "beginning," and "strength" rather than bovine contexts, reflecting its Egyptological function as an ideogram for *ḫꜥt* "forepart" (Gardiner, 1957, p. 461). Each entry records the sign's semantic meaning, Gardiner category, usage context description, and a set of keywords for semantic field matching. The knowledge base also maintains two auxiliary structures: a set of signs that are *never* determinatives (e.g., common phonograms such as G17, N35, S29) and a five-tier sign categorisation system.

The five-tier categorisation system assigns each sign a base score reflecting its likelihood of functioning as a determinative. Tier 1 (`phonogram_strict`) receives a score of 0, representing signs that never function as determinatives. Tier 2 (`phonogram_primary`) receives a score below 0.1. Tier 3 (`context`) is subject to a context-dependent threshold gate. Tier 4 (`reliable`) receives a bonus of 0.10–0.12. Tier 5 (`plural_functional`) receives a score of 0.05, covering plural markers such as Z2 and Z3.

3.6 Word Boundary Agent

The Word Boundary Agent is a deterministic (non-LLM) processing node that performs three critical functions. First, it infers word and compound boundaries within the MdC sequence, treating the entire sequence as a single compound word when explicit word delimiters are absent. Second, the agent generates *contiguous suffix span candidates* for each identified word. This step implements the well-established Egyptological principle that determinatives occupy word-final positions and form contiguous blocks (Goldwasser, 2002; Goldwasser and Grinevald, 2012). For each word, a tail window (typically the last 4 or 6 tokens) is examined, and all valid contiguous suffix spans are enumerated, including the empty span (representing "no determinative for this word"). Third, the agent computes *phonetic complement hints* by identifying common uniliteral phonograms (e.g., M17 */i/*, D21 */r/*, G17 */m/*, N35 */n/*) appearing near word endings, signs that frequently serve as pronunciation aids and should not be classified as determinatives by default.

3.7 Agent 1: Semantic Analysis

The semantic agent is a large language model (configurable as OpenAI GPT or Anthropic Claude) tasked with identifying determinatives through semantic and cultural reasoning. The agent receives a carefully engineered system prompt encoding Egyptological principles and domain-specific rules. A critical component of the prompt engineering strategy is the inclusion of context-dependent sign rules for signs with dual function (phonogram versus determinative). For instance, vessel/container signs (W22, W24, W23) are validated as determinatives only when vessel, liquid, or offering semantic content is present; divine/sacred signs (R8, R7) require divine or sacred context; and animal signs (I9, I10) are treated as phonograms by default unless overwhelming reptile context is present.

The agent's input includes the MdC sequence, the indexed token list, candidate information from the knowledge base, phonetic complement hints, structural hints (block finals, groups), and the rule-based baseline. The agent returns a structured JSON containing identified determinatives (each with token name, zero-based index, and confidence), functional markers, a natural-language justification, and alternative interpretations. The prompt engineering strategy draws on chain-of-thought reasoning (Wei et al., 2022) to elicit structured analytical outputs. This output is stored in the shared state as `semantic_json` and is consumed by the Synthesiser Agent but *not* by the Structural Agent, preserving agent independence.

3.8 Agent 2: Structural Analysis (Independent)

The structural agent analyses the same input sequence from a complementary perspective, focusing exclusively on positional and distributional patterns rather than semantic content. This agent identifies determinatives based on three primary structural criteria: (1) position within hieroglyphic blocks, (2) plural marker patterns (Z1, Z2, Z3 clusters), and (3) stacking and ligature patterns suggesting functional grouping.

A key architectural decision is the complete independence of the structural agent from the semantic agent's output. This independence is essential for producing uncorrelated evidence streams (Wang et al., 2023): when both agents independently identify the same sign as a determinative, their agreement constitutes stronger evidence than either agent's assessment alone. An earlier sequential architecture, in which agents processed the input in series and could observe each other's outputs, was tested but produced inferior results, as downstream agents tended to anchor on upstream decisions rather than forming independent judgements.

3.9 Agent 3: Synthesiser

The synthesiser agent serves as the integration point for multi-source evidence, operating after both analysis agents have completed their independent assessments. This agent receives the MdC sequence, all preprocessing information, and both the semantic and structural JSON outputs. Its role is to perform evidence triangulation, examining where the two agents agree, identifying contradictions, and proposing a consolidated interpretation that weighs the relative strength of semantic versus structural evidence.

The synthesiser produces a comprehensive output containing four components. First, the *main interpretation* provides a consolidated list of determinatives with each sign annotated by its index, Gardiner code, meaning, category, and identification method. Second, up to three *alternative interpretations* are generated, each labelled with a perspective and a *promotion condition*, the minimal additional evidence required for this alternative to supersede the main interpretation. Third, an *ambiguity analysis* summarises interpretive uncertainty. Fourth, *recommendations* provide actionable guidance for resolving remaining ambiguities.

3.10 Deterministic Scoring Module

The deterministic scoring module post-processes the synthesiser's output through five sequential sub-stages, applying hard constraints and quantitative scoring to refine

LLM-generated classifications. The numerical parameters governing the scoring module, including source weights, context thresholds, positional bonuses, and penalty values were determined empirically through iterative experimentation on a development subset of annotated sequences, optimising for balanced precision and recall. These values represent the configuration that yielded the best overall performance and should be understood as empirically calibrated rather than theoretically derived. The five sub-stages are as follows:

- *Evidence aggregation* collects candidate determinatives from three sources and computes a per-source confidence score: 38% weight for semantic evidence, 34% for structural evidence, and 28% for rule-based evidence. Multi-source agreement triggers a bonus (+8% for two-source agreement, +6% for three-source agreement), rewarding convergent evidence.
- *Sign category validation* applies the five-tier categorisation system described in Section 3, adjusting each candidate's score according to its tier assignment. Signs in Tier 1 receive a score of zero regardless of agent evidence, effectively implementing a hard veto. Signs in Tier 3 pass through a threshold gate requiring additional semantic validation.
- *Semantic context validation* applies ten semantic field categories with per-sign thresholds. For example, the sign I9 (horned viper) requires a threshold of 0.90, reflecting its overwhelmingly phonographic function, while W24 (vessel) requires 0.75 and R8 (divine sign) requires 0.70. If a sign's context score falls below its threshold, it is rejected as a determinative regardless of agent consensus.
- A *recently-added penalty* applies proportional score reductions to signs recently added to the knowledge base. Signs I9 and I10 receive a 40% penalty, D58 and U1 receive a 25% penalty, and other recently added signs receive a 15% penalty.
- Finally, *span ranking and selection* operates on the contiguous suffix spans generated by the Word Boundary Agent. For each word, the module scores every valid suffix span by aggregating token-level scores, applying a phonetic complement penalty (-0.15 per phonetic complement within the span), and adding positional bonuses ($+0.10$ for word-final positions, $+0.06$ for near-final positions). The highest-scoring span is selected as the main interpretation, with up to three alternative spans retained.

Importantly, span scoring operates on each sign independently; co-occurring signs from the same Gardiner section (e.g., O31, the door leaf used as a determinative for "door" and "opening," and O49, the town/city determinative) are not assumed to exhibit semantic synergy, as their determinative functions are Egyptologically distinct (Gardiner, 1957, pp. 496–498). Each sign must independently satisfy its context-dependent threshold to be retained in the final span.

3.11 Structured Output

The system produces a structured output for each input sequence comprising four sections: (1) the *Main Interpretation*, presenting the primary determinative set with sign-by-sign justifications; (2) the *Alternatives* section, providing up to three ranked

alternative interpretations, each with an explicit promotion condition; (3) the *Ambiguity Analysis* section, cataloguing factors contributing to interpretive uncertainty; and (4) the *Recommendations* section, offering actionable guidance for resolving remaining ambiguities. By providing ranked hypotheses with explicit promotion conditions, the system acknowledges the interpretive multiplicity inherent in Egyptological analysis while still offering actionable guidance.

3.12 Evaluation Protocol

The system was evaluated on a benchmark of 99 hieroglyphic sequences annotated by expert Egyptologists. Rather than relying on a single development/test partition, which leaves test performance vulnerable to sampling chance and provides no estimate of variance, we adopted a *nested cross-validation* protocol that disentangles parameter selection from generalisation estimation (Cawley and Talbot, 2010; Varma and Simon, 2006). This design serves two purposes: (i) every sequence in the corpus contributes to the final performance estimate as a held-out test case exactly once, and (ii) the variance across folds quantifies the robustness of the system to changes in the training distribution.

Outer loop to generalisation estimation.

The 99 sequences were partitioned into $K_{\text{outer}} = 5$ disjoint folds via a stratified shuffle with a fixed random seed for full reproducibility. Each outer iteration removes one fold (19–20 sequences) as a test partition and uses the remaining 79–80 sequences for parameter selection. The five outer iterations cover the entire corpus, so the aggregate test performance is computed across all 99 sequences.

Inner loop to parameter selection.

On each outer training partition we run a Dietterich 5 $\times$2 cross-validation (Dietterich, 1998) to optimise the parameters of the deterministic scoring module. The training partition is split into two equal halves, and each half serves once as the optimisation half and once as the validation half. This is repeated five times with different random splits, yielding 10 validation estimates per outer fold. Within each repetition the optimisation half is searched by a Bayesian sampler (Tree-structured Parzen Estimator; Bergstra et al., 2011) implemented in Optuna (Akiba et al., 2019) over the parameter space of context thresholds, source weights, agreement bonuses and category-specific penalties, with macro-F1 on the validation half as the objective.

Test inference.

After the inner loop yields a set of optimised parameters for each outer fold, the full multi-agent pipeline, including the three LLM agents and the deterministic scoring module, is executed on the held-out test fold with those parameters. Test predictions are therefore generated by the complete system under the same parameter selection discipline used for training, but on sequences that played no role in either parameter optimisation or method development. All Anthropic LLM calls are issued with `temperature` = 0 to maximise output stability across runs. The inner-loop random

state, the outer-fold shuffle, the Optuna sampler and all NumPy/Python random generators are seeded from a single global constant.

Metrics.

For each test fold we compute macro- and micro-averaged precision, recall and F1, treating empty predictions on empty ground truths as correct nulls so as not to inflate macro-F1 with an undefined denominator. The four sequence-level outcome categories (STANDARD, TRUE NULL, FP NULL, FN NULL) are tracked separately so that the system's behaviour on edge cases is not hidden by token-level averaging. Across folds we report the mean and standard deviation; the standard deviation is the primary indicator of robustness because it captures how much the absolute performance depends on the particular partition.

3.13 Use of Large Language Models

Large language models are a component of the method itself: the semantic, structural, and synthesiser agents described earlier in this section are LLM-based, and their configuration, prompting strategy, and inference settings are documented there. Separately, in preparing this manuscript the authors used an AI assistant to translate portions of the text from Portuguese into English and to assist with the synthesis of the reviewed literature. All resulting content was reviewed and edited by the authors, who take full responsibility for the final version of the manuscript.

4 Results

4.1 Overall Performance under Nested Cross-Validation

We report the performance of the proposed method under the nested cross-validation protocol described in Section 3.12. All metrics below are computed on held-out test folds; every one of the 99 corpus sequences serves as a held-out test case exactly once.

Across the five outer folds, the system achieves a macro-averaged F1 of 81.94% ±2.19% and a micro-averaged F1 of 87.50 % ± 2.50 %, with macro-precision of 87.20% ±2.78% and macro-recall of 86.72% ±5.07% (Table 1). The corresponding micro-averaged precision and recall, computed per fold from the token-level TP/FP/FN totals, are 88.19% ±0.51% and 87.01% ±5.19% respectively. Figure 2 shows the fold-level distribution of each of these metrics side by side: micro variants are consistently higher than their macro counterparts and have lower spread, reflecting the dominance of high-frequency common signs in the corpus.

The small standard deviation across folds (2.19 pp for macro-F1, and a particularly tight 0.51 pp for micro-precision) is informative on its own: it indicates that the system's performance is not the product of a fortunate train/test split. The inner-CV validation F1 averaged 90.42% ±0.87% across the same five outer training partitions, again with low variance, suggesting that the optimised parameters themselves are stable across resamplings of the corpus.

Table 1 Held-out test performance of the proposed method under nested 5-fold outer cross-validation with inner 5 ×2 CV parameter optimisation. All values are means across the five outer folds; standard deviation in parentheses. Each of the 99 corpus sequences serves as a held-out test sample exactly once. Micro precision and recall are computed per fold from the token-level TP/FP/FN totals of that fold's test set, then averaged across folds.

Metric	Macro (%)	Micro (%)
F1-Score	0.8194 (0.0219)	0.8750 (0.0250)
Precision	0.8720 (0.0278)	0.8819 (0.0051)
Recall	0.8672 (0.0507)	0.8701 (0.0519)

Table 2 Per-fold test performance of the proposed method. n_{train} is the size of the outer training partition (used for inner-CV parameter selection); n_{test} is the size of the held-out test partition. "Val F1 (mean±std)" is the inner-CV validation F1 averaged over the 5 ×2 inner repetitions.

Fold	n_{train}	n_{test}	Val F1 (mean±std)	Macro F1	Micro F1	Macro P	Macro R
1	79	20	0.9205 ± 0.0122	0.7762	0.8296	0.8633	0.7925
2	79	20	0.8978 ± 0.0174	0.8325	0.8730	0.9199	0.8375
3	79	20	0.8973 ± 0.0119	0.8354	0.9009	0.8667	0.8875
4	79	20	0.9056 ± 0.0063	0.8269	0.8947	0.8340	0.9442
5	80	19	0.8998 ± 0.0236	0.8260	0.8769	0.8760	0.8743
Mean	—	—	0.9042 ± 0.0087	**0.8194**	**0.8750**	**0.8720**	**0.8672**
Std	—	—	—	0.0219	0.0250	0.0278	0.0507

Per-fold breakdown.

Table 2 presents the per-fold performance of the full system. The macro-F1 range across folds is 77.62%–83.54% (spread of 5.92 pp), tighter than would be expected from the variability typically observed in corpora of this size. Four of the five folds fall within 2 pp of the mean (82.60%–83.54%); the only outlier is fold 1, where the test partition concentrates several rare-sign sequences including the corpus-unique `US7063L36VARA` variant. Even on this adversarial fold the macro-F1 of 77.62% remains above 75%, and the macro-precision of 86.33% is in line with the overall trend.

Aggregated token-level confusion.

Pooling token-level decisions across all five test folds yields 269 true positives, 36 false positives and 42 false negatives over 99 sequences, giving an aggregate token-level micro precision of 88.20%, recall of 86.50% and F1 of 87.34% (Table 3). Critically, every corpus sequence contributes to these totals as a test case, so this aggregate is

Table 3 Token-level confusion totals aggregated across all five held-out test folds. Each of the 99 corpus sequences contributes to this matrix as a test sample exactly once.

	Pred. Positive	**Pred. Negative**	**Total**
Actual Positive	269 (TP)	42 (FN)	311
Actual Negative	36 (FP)	—	—

not subject to the train/test selection bias that a single-split confusion matrix would carry.

4.2 Generalisation: Validation vs. Test

Figure 3 plots the inner-CV validation F1 for each outer fold against the corresponding test-set macro-F1. All five fold points lie below the diagonal, with a mean gap of -8.5 pp between validation (90.4%) and test (81.9%) macro-F1. This gap is the cost of two factors that the inner CV cannot model: the LLM inference path (the inner loop optimises rule parameters using the rule-only path) and the unseen-sign incidence in each outer test fold. The gap is highly consistent across folds (std 3.0 pp), which is what allows the test-set estimate to be trusted: were the gap dominated by random variability, its standard deviation would stunt its mean.

4.3 Sequence-Level Outcome Distribution

Each held-out test prediction is categorised into one of four sequence-level outcomes to expose behaviour that token-level averaging would hide:

- Standard: both prediction and ground truth contain at least one determinative (the dominant case);
- True Null: prediction and ground truth are both empty (correctly abstaining);
- FP Null: prediction contains signs but the ground truth is empty (spurious detection);
- FN Null: prediction is empty but the ground truth contains determinatives (missed detection).

Aggregated across all five test folds (99 total sequences), the breakdown is 88 Standard, 5 True Null, 4 FP Null and 2 FN Null. Among the 88 Standard sequences, 40 (45.5 %) achieve perfect token-level F1, 87 (98.9 %) achieve F1 ≥ 0.5, and only 1 sequence falls below F1 = 0.5: the corpus-unique `US7063L36VARA` case mentioned above. The median per-sequence F1 on Standard cases is 0.889, with a mean of 0.865.

The four FP Null cases (false positives on empty ground truth) are concentrated in folds 1, 4 and 5 and involve signs T18, X3, D4+A1 and W24. The two FN Null cases (missed detections) appear in folds 2 and 3 and involve Aa3 and T14, both rare forms with low knowledge-base coverage. Figure 4 shows the per-fold distribution of these outcomes.

4.4 Performance by Determinative Category

Beyond aggregate token-level metrics, the cross-validated nature of the evaluation makes it possible to break down performance by the *semantic category* of the determinative being predicted, attributes such as PERSON, ANIMAL, DIVINE, NATURE, BODY, TOOL, OBJECT, ACTION and so on. Because every sequence appears in exactly one held-out test fold, the pooled per-category precision, recall and F1 are computed over the entire corpus and are not subject to train/test selection bias. Figure 5 reports these values for every category attested at least three times in the ground truth.

Cluster into three parts

. The first part containing FUNERAL, GRAMMAR, NATURE, SKY, TOOL, PERSON and DIVINE achieves F1 ≥ 0.95 with precision close to 1.00 and recall above 0.95. These categories are well covered by the knowledge base and have distinctive structural cues (FUNERAL signs occupy near-terminal positions; DIVINE signs combine with strong disambiguators such as A40 and R8). The second part contains mid-range categories (ABSTRACT, BODY, ACTION, LAND, FOOD, TRANSPORT) with F1 between 0.66 and 0.94: these are categories where the LLM agents and the rule engine occasionally disagree, and the deterministic scoring module is the arbiter. The third ´part contains ANIMAL and OBJECT, the two outliers: precision drops to 0.10–0.25 while recall is much higher. The signature is that of a system over-proposing candidates within these categories, a useful diagnostic, because it points directly at the part of the agent guidance that warrants revision in future work.

A second observation, consistent with what was previously described qualitatively as the "complexity paradox", is that high-complexity sequences (those containing three or more determinatives) remain by far the easiest case for the system. They constitute the large majority of the corpus, and they are also where independent evidence streams reinforce one another, the conditions under which a multi-agent architecture is expected to excel. Sparse sequences, particularly those with no determinatives, are where the absence of positive evidence makes spurious detections most damaging (Section 4.3).

4.5 Per-Sign Error Distribution

Because every corpus sequence appears in exactly one held-out test fold, the per-sign F1 score can be computed by pooling true positives, false positives and false negatives across all five folds and grouping them by sign code. This yields an unbiased per-sign assessment over the full corpus. Figure 6 reports the per-sign F1 for all signs with at least three occurrences in the ground truth, coloured by category.

The high-frequency core of the corpus is detected with near-perfect F1. A1 ($n = 55$), Z1 ($n = 49$), Z2 ($n = 35$), A40 ($n = 23$) and N1 ($n = 13$), collectively responsible for 56% of all determinatives in the corpus, all achieve F1 ≥ 0.94. The seventeen signs with F1 = 1.00 (perfect detection across all folds in which they appear) span every major category: PERSON, LAND, TOOL, NATURE, DIVINE, ABSTRACT, BODY, ACTION, GRAMMAR and UNKNOWN, showing that the architecture's strengths are not confined to a single domain.

The remaining error mass is concentrated on two diagnostically revealing sign types. D54 (walking legs; action category) reaches only F1 = 0.60 across 12 ground-truth occurrences: it is a high-frequency dual-function sign, behaving as a motion determinative in some contexts and as a verbal ideogram in others, and the deterministic scoring module currently lacks the lexical resolution to draw that distinction reliably. US7063L36VARA (F1 = 0) is a corpus-unique non-standard notation absent from the Gardiner-derived knowledge base; it is a data-coverage limitation rather than a methodological one, addressable by knowledge-base expansion.

This per-sign pattern is qualitatively different from what a fixed train/test split could have revealed: under cross-validation the estimate of each sign's F1 draws on every occurrence in the corpus, which means rare signs are not arbitrarily classified as "never evaluated" because they happened to fall on the wrong side of a single partition.

4.6 Illustrative Outputs

To illustrate the system's output format and the richness of its analytical products, three representative examples are presented below, selected to demonstrate different aspects of system behaviour: a high-confidence correct prediction, a case where the correct answer appears in the alternatives, and a case illustrating context-dependent sign rejection.

4.6.1 Example 1: High-Confidence Multi-Determinative Sequence (Perfect Match).

Input: `W24-V31-A1-W24-X1-N1-W25-N35-N35-V31-O93-X1-B1-D46-X1-T28-D21-D1-Z1-V31-N35-M36-S29-D36-Y2-V31-D4-Q1-A40-R16-S29-T18-A1-D38-S29-N35-V31-F34-Z1-V31-G17-F32-X1-Z1-V31-D4-Q1-A40-R16-S29-T18-A1`

Listing 1 Structured output for Example 1.

```
MAIN INTERPRETATION:
Determinatives: [A1, N1, B1, Z1, A40, A1, Z1, Z1, A40, A1]
Justification:
  A1 (pos 2) : masculine person (semantic) + word-final (structural)
  N1 (pos 5) : sky/heaven (semantic) + word-final (structural)
  B1 (pos 12): woman (semantic) + word-final (structural)
  Z1 (pos 18): singular marker (semantic) + functional (structural)
  A40 (pos 28): divine person (semantic) + word-final (structural)
  A1 (pos 32): masculine person (semantic) + word-final (structural)
  Z1 (pos 38): singular marker + functional marker
  Z1 (pos 42): singular marker + functional marker
  A40 (pos 46): divine person (semantic) + word-final (structural)
  A1 (pos 50): masculine person (semantic) + word-final (structural)
Semantic interpretation: Multi-word sequence referencing masculine,
  feminine, and divine persons with singular markers.
Confidence: HIGH

ALTERNATIVE INTERPRETATION 1:
  Perspective: Structural (functionalist)
  Additional determinatives: [Y2 (pos 24), D36 (pos 23)]
  Promotion condition: If epigraphic parallels show D36/Y2 functioning
    as classifiers in offering-text contexts, promote to determinative.
```

```
AMBIGUITY ANALYSIS:
  Degree: LOW
  Factors: Strong inter-agent consensus on all 10 determinatives.
    Minor ambiguity on Y2 and D36 (phonogram vs. determinative).

RECOMMENDATIONS:
  Consult parallel inscriptions for D36/Y2 classification in
  offering formulae. Main interpretation follows conservative criteria.
```

This example demonstrates the system handling a long sequence (52 tokens) containing 10 determinatives across multiple words. The sign A1 appears three times, Z1 three times, and A40 twice, each listed separately with its specific position. The alternative interpretation suggests two additional candidates from a functionalist perspective, with an explicit promotion condition specifying what evidence would resolve the ambiguity. The system correctly matched the expert annotation (TP = 10, FP = 0, FN = 0, F1 = 100%).

4.6.2 Example 2: Partial Match with Correct Answer in Alternatives.

Consider a sequence where the expert identified two determinatives but the system's main interpretation included only one, while the second appeared in an alternative. This pattern was observed in 42% of cases where alternatives contained correct or superior solutions.

Listing 2 Structured output for Example 2.

```
MAIN INTERPRETATION:
  Determinatives: [O39]
  Justification:
    O39 (pos 4): building/construction (semantic + structural).
    Both semantic and structural agents agree on this classification.
  Confidence: HIGH

ALTERNATIVE INTERPRETATION 1:
  Perspective: Functionalist school
  Determinatives: [penultimate sign, O39]
  Promotion condition: If the penultimate sign has no clear phonetic
    value in this context AND parallel texts show it functioning
    as a form/object classifier, include as second determinative.

AMBIGUITY ANALYSIS:
  Degree: MEDIUM
  Factors: Polivalence of penultimate sign (phonogram vs. determinative).
    Different Egyptological schools would classify differently.
    Interpretation A (1 det.): traditional positional criterion.
    Interpretation B (2 dets.): functionalist/semantic.
    No universal Egyptological consensus on this specific case.
```

In this case, the expert annotated two determinatives, matching the alternative interpretation. The system's main interpretation achieved F1 = 66.7% (1 TP, 0 FP, 1 FN), but by considering the first alternative, the correct answer was available. The ambiguity analysis explicitly identifies the polivalence of the penultimate sign and acknowledges that different Egyptological schools would classify the sequence differently, transforming a partial "error" into a documented interpretive choice.

4.6.3 Example 3: Context-Dependent Sign Rejection.

This example illustrates how the semantic context validation mechanism prevents false positives by rejecting signs that lack appropriate semantic context.

Listing 3 Structured output for Example 3.

```
MAIN INTERPRETATION:
  Determinatives: [A1, D54, Z3]
  Justification:
    A1 (pos 15): masculine person (semantic) + word-final (structural)
    D54 (pos 8): motion/walking (semantic) + word-final (structural)
    Z3 (pos 16): plural marker (structural pattern)
  Signs considered but REJECTED:
    W24 (pos 0): vessel sign rejected -- no vessel/liquid/offering
      semantic content. Context: human actions. Failed threshold (0.75).
    R8 (pos 3): divine sign rejected -- no divine/sacred semantic
      content. Secular context. Failed threshold (0.75).
  Confidence: HIGH

ALTERNATIVE INTERPRETATION 1:
  Perspective: Expanded semantic
  Determinatives: [W24, A1, D54, Z3]
  Promotion condition: If broader text context reveals an offering
    scene, revalidate W24 as vessel determinative.

RECOMMENDATIONS:
  Verify whether broader textual context supports vessel/offering
  reading. If so, W24 should be promoted per Alternative 1.
```

The signs W24 (vessel) and R8 (divine) were identified as candidates by both agents but rejected by the deterministic scoring module because the surrounding semantic context (human actions, walking, speaking) did not meet the required thresholds (0.75 for both signs). The system documents these rejections transparently, explaining why each sign was excluded and under what conditions it would be reinstated. This level of transparency transforms the scoring module's hard constraints into interpretable decisions.

4.7 Value of Alternative Interpretations

Across the 99-sequence corpus, the system generated a mean of 2.9 alternatives per sequence, with 89.9% receiving at least one alternative and 68.7% receiving the maximum three. Manual review of a stratified sample of 30 sequences revealed that 42% of alternative interpretations contained the correct determinative set or a better approximation than the main interpretation, while 31% contained partially correct alternatives that, if combined with the main interpretation, would yield a complete solution. This pattern held consistently across all five outer test folds, confirming that the multi-hypothesis generation capability is not an artefact of parameter calibration but a genuine architectural property. Common promotion conditions referenced epigraphic parallels, semantic field verification, and grammatical disambiguation.

5 Discussion

5.1 Performance in Context

Under nested cross-validation, the proposed method achieves a macro-F1 of 81.94% ±2.19% and a micro-F1 of 87.50% ±2.50% on held-out test folds covering the entire 99-sequence corpus. To our knowledge, no previously published system addresses automatic determinative detection in Egyptian hieroglyphic sequences, so the proposed method establishes a first reference point on this task. The macro-precision (87.20%) and macro-recall (86.72%) are nearly balanced, indicating that the system is neither over- nor under-classifying on aggregate, a result of the deterministic scoring module being calibrated by the inner CV rather than tuned by hand on a single split.

The choice of evaluation protocol matters here. A single train/test split yields a single point estimate whose distance from the true expected performance is unknown. Thenested CV design produces five independent point estimates whose standard deviation (2.19 pp for macro-F1) directly quantifies the robustness of the system to changes in the training set. The variance is small enough that the headline metrics can be reported with confidence; it is also large enough to make explicit that the system is not equally accurate on every conceivable corpus partition.

The micro-F1 is consistently 5–6 pp higher than the macro-F1 across folds. This well-known asymmetry reflects the corpus class distribution: common signs like A1, Z1, Z2 and A40 dominate the token totals (and hence the micro-average) and are detected at near-perfect F1, while rare signs with low support drag the macro-average down. The micro-F1 is the more relevant metric for downstream applications such as auto-annotation of large corpora; the macro-F1 is the more demanding methodological benchmark.

5.2 The Complexity Paradox

The earlier observation that performance improves with sequence complexity is reinforced rather than contradicted by the cross-validated evaluation. Across all five test folds, sequences with three or more determinatives remain the easiest case, and empty sequences, which require correct abstention rather than positive detection, remain the hardest. Of the 11 sequences in the corpus whose ground truth is either empty or single-determinative, six produce a non-STANDARD outcome (5 TRUE NULL and 4 FP NULL cases combined), so the empty-sequence challenge accounts for an outsized share of residual error.

We continue to attribute this to multi-agent triangulation. In information-rich sequences the semantic agent accesses richer contextual cues, the structural agent identifies more reliable positional patterns, and the rule engine provides more candidate matches. Evidence triangulation naturally produces higher confidence when convergent evidence is abundant. Conversely, in sparse sequences the system becomes more susceptible to noise, the FP Null cases involving T18, X3, D4+A1 and W24 (Section 4.3) are the visible manifestation of this. A dedicated empty-sequence detector remains the most direct route to closing this gap and is discussed under future work.

5.3 Error Analysis and Dual-Function Signs

The pooled per-sign analysis (Section 4.5, Figure 6) and the per-category breakdown (Figure 5) provide a more reliable picture of error concentration than any single train/test split could. Two sign types stand out as the dominant residual error source: D54 (walking legs, F1 = 0.60 across 12 ground-truth occurrences) and the corpus-unique `US7063L36VARA` variant (F1 = 0 across 3 occurrences). These two account for the bulk of the gap between the near-perfect detection of common signs (A1, Z1, Z2, A40, N1, all ≥ 0.94 F1) and the macro-averaged 81.94%.

D54 is the most diagnostically significant case: it is a high-frequency dual-function sign, behaving as a motion determinative in some contexts and as a verbal ideogram in others. Its misclassification in both directions (false positives on phonographic occurrences and false negatives on determinative ones) indicates inconsistency rather than systematically biased detection. The system has learned appropriate sensitivity, i.e., it recognises D54's potential determinative function, but lacks the lexical discrimination to decide reliably when that function is activated. A1 (seated man) shows a milder version of the same pattern: oscillating between personal classifier and phonographic use, its classification depends on whether the host word is a substantive or a verb, a distinction only partially captured by the current semantic field validation mechanism.

US7063L36VARA is a different kind of failure: a knowledge-base coverage gap rather than a contextual ambiguity. The variant notation does not appear in the Gardiner-derived sign inventory, so the rule engine cannot recognise it and the LLM agents fall back on phonographic interpretation. This is the clearest example of where corpus-derived knowledge-base expansion would have a direct effect on system performance. The ANIMAL and OBJECT category outliers in Figure 5 represent a milder coverage issue, where the agents over-propose category-plausible candidates that are not present in the ground truth.

5.4 Generalisation Analysis

Under the nested CV protocol, generalisation is no longer a single dev-vs-test comparison but a structural property of the evaluation: every sequence in the corpus contributes to the test estimate as a held-out case exactly once. Three observations from Section 4.1 bear on the generalisation claim.

- **Cross-fold stability.** The macro-F1 standard deviation across the five outer folds is 2.19 pp (Table 1), small relative to the 81.94% mean. Of the five folds, four lie within 2 pp of the mean (82.6%–83.5%); fold 1 is the only outlier at 77.6% and concentrates several rare-sign sequences including the corpus-unique `US7063L36VARA` notation. This variability is consistent with a system whose dominant failure modes are tied to specific sign types rather than to spurious sensitivity to the articular partition.
- **The validation–test gap.** The mean gap of -8.5 pp between inner-CV validation F1 (90.4%) and held-out test F1 (81.9%) is informative rather than alarming (Figure 3). It has a coherent explanation: the inner CV optimises the deterministic scoring module using the rule-only path (no LLM inference), so it cannot account for the behaviour of the LLM agents at test time. The gap is consistent across folds (std 3.0 pp), which rules out the possibility that it is driven by occasional outlier

folds. In effect, the inner CV gives an optimistic but stable upper bound for what the full system can achieve; the test set establishes the realised performance.

- **Inner-CV convergence.** Within each outer fold, the 5 $\times$2 inner CV converges quickly and consistently: validation F1 standard deviation across the ten inner repetitions averages 0.014 across all outer folds (Table 2), and never exceeds 0.024. The inner CV therefore does not introduce additional variance to the test estimate; the test variance comes from the inherent difficulty of the outer fold itself.

Together, these observations indicate that the proposed architecture generalises in the sense the term carries under cross-validation: performance does not depend on a particular held-out partition, the sources of variation are structurally identified, and the absolute test performance can be reported with a meaningful confidence quantification. The point estimates obtained on any single train/test split should therefore be read against the $\pm$2.19 pp macro-F1 envelope established here, rather than as exact values.

5.5 Alternative Outputs as a Methodological Innovation

The generation of ranked alternative interpretations with explicit promotion conditions constitutes one of the main contributions of this work. The system operates as a ranking algorithm, surfacing the k most probable interpretations ($k \leq 4$: one main plus three alternatives) and allowing the scholar to select based on contextual knowledge the system lacks. As Example 2 in Section 4 demonstrated, a sequence classified as a "partial error" under traditional evaluation actually reflects a legitimate interpretive choice, with the alternative hypothesis correctly documenting the competing reading.

5.6 Multi-Agent Architecture: Strengths and Implications

The independence of the semantic and structural agents ensures that their agreement constitutes genuine evidential convergence rather than shared bias. This is analogous to inter-rater reliability in manual annotation (Cohen, 1960): agreement between independent raters carries more weight than agreement between raters who have discussed their judgements.

The hybrid architecture, combining LLM-based contextual reasoning with deterministic rule-based constraints, addresses a fundamental tension in applying language models to specialised domains. LLMs provide flexible contextual understanding and can reason about semantic fields, cultural associations, and unusual patterns that rigid rules would miss (Yao et al., 2023). However, LLMs can also hallucinate, produce inconsistent outputs, and fail to respect domain-specific hard constraints. The deterministic scoring module serves as a guardrail, enforcing tier-based restrictions, semantic context thresholds, and structural validity constraints that the LLMs alone cannot guarantee. The conservative expansion strategy, implemented through the recently-added penalty, addresses the challenge of growing the knowledge base without inflating false positives.

5.7 Explainability and Scholarly Utility

This level of transparency is essential for adoption in Egyptological scholarship, where computational tools are viewed with appropriate scepticism (Assael et al., 2022) and scholars require not merely predictions but reasoning they can evaluate and contest. The structured output format transforms the system from a black-box classifier into a collaborative tool, complementary to digital annotation platforms such as iClassifier (Harel et al., 2024), that surfaces analytical possibilities and frames interpretive decisions in terms familiar to domain experts.

5.8 Limitations and Future Work

Several limitations merit discussion. The evaluation corpus of 99 sequences is relatively small for drawing strong generalisation claims beyond Middle Egyptian funerary and administrative texts from the TLA (Thesaurus Linguae Aegyptiae, 2024). Performance on Old Kingdom, Late Period, or demotic texts, which exhibit different orthographic conventions and sign inventories, remains untested (cf. Lincke and Kammerzell, 2012).

The pooled outcome breakdown (Section 4.3) shows that of the 9 sequences with empty ground truth in the corpus, 5 were correctly classified as TRUE NULL and 4 produced FP NULL outcomes. Two additional FN NULL cases involve sequences where the system produced no prediction despite a non-empty ground truth. Together, these 6 null-related errors account for a disproportionate share of residual misclassification, especially on macro-averaged metrics where rare-sign folds are weighted equally. A dedicated empty-sequence detector would address the FP NULL cases directly. For single-determinative sequences, enhanced contextual modelling through sequence-level embeddings or full-sequence attention could provide the discriminative power currently lacking. The cross-validated evaluation makes it possible to quantify ex ante the macro-F1 lift such a detector would deliver: replacing the four FP NULL predictions with correct nulls would close roughly 4 pp of the gap between macro-F1 (81.94%) and micro-F1 (87.50%).

The pooled per-sign analysis (Section 4.5) identifies two concentrated sources of residual error that the architecture itself cannot resolve. D54 (walking legs) reaches only F1 = 0.60 across its twelve occurrences and is the most diagnostically significant case: its dual function as motion determinative and verbal ideogram requires lexical-level discrimination that the current scoring module does not encode. The corpus-unique `US7063L36VARA` variant (F1 = 0) is a straightforward knowledge-base coverage gap, addressable by incorporating non-standard variant notations into the sign inventory. The per-category analysis (Section 4.4) additionally flags ANIMAL and OBJECT categories as having unusually low precision, indicating that the LLM agents over-propose category-plausible candidates in these domains; targeted revision of the agent guidance for these two categories is a low-effort, well-localised intervention that should improve macro-F1 without touching the scoring module.

The inner 5 $\times$2 CV currently optimises only the deterministic scoring module's parameters, using the rule-only inference path as its training signal. This choice was deliberate: rule-only inference is fast enough to support thousands of Bayesian evaluations per outer fold, whereas full LLM inference is not. The consequence, visible

in Figure 3, is a consistent -8.5 pp gap between inner-CV validation F1 and held-out test F1, attributable to the unmodelled LLM inference path rather than to overfitting. Closing this gap would require an LLM-aware inner loop, for example, optimising agent guidance prompts or LLM-specific aggregation weights under a constrained budget of full pipeline evaluations. Caching mechanisms for repeated LLM calls on the same input could make such an extension computationally tractable.

A further extension would involve processing sequences within their broader textual context. Since the system already generates recommendations specifying what additional evidence would resolve ambiguities, consecutive sequences from the same inscription could provide precisely this context, potentially addressing errors in sparse sequences where intra-sequence evidence is insufficient.

Each test sequence requires three LLM calls with combined processing time of approximately 10–16 seconds. The current implementation issues all LLM calls at `temperature` = 0 to maximise output stability; the Anthropic API does not currently expose a `seed` parameter, so strict bit-for-bit reproducibility of LLM outputs is not guaranteed, although the variance contribution of LLM sampling to fold-level metrics is small (visible in the tight ± 0.51 pp standard deviation of micro-precision across folds). Future work could explore distillation of agent knowledge into smaller, faster local models, which would simultaneously reduce latency and restore full determinism.

The current scoring module parameters are selected by Bayesian optimisation against an empirical macro-F1 objective. A complementary approach would derive parameters from corpus-derived frequency statistics. Similarly, the source weights for evidence aggregation could be grounded in information-theoretic measures or formal evidence combination frameworks such as Dempster–Shafer theory (Shafer, 1976). The system's structured output format also creates a natural pathway for expert-in-the-loop learning, where scholarly corrections on prior outputs could be incorporated to refine knowledge base entries between cross-validation rounds.

The current evaluation treats all determinative errors equally. A false positive Z1 may be far less consequential than a false negative A40 (male divine classifier) for downstream interpretive use. Future evaluation frameworks could incorporate weighted metrics reflecting the differential importance of determinative types for semantic disambiguation, and could be plugged into the same nested CV infrastructure without methodological change.

Finally, the nested cross-validation as implemented uses five outer folds with a single random partition; the inner 5×2 CV is internally repeated but the outer partition is not. Repeating the outer partition under several random seeds and pooling the resulting test estimates would yield narrower confidence intervals on all reported metrics. With the current configuration this would multiply the computational cost of evaluation linearly, but it is a straightforward extension and would strengthen the statistical claims that can be made on a corpus of this size.

6 Conclusion

This paper presented a system, a multi-agent framework for automatic determinative detection in Egyptian hieroglyphic sequences encoded in MdC notation, a task

that, to our knowledge, has not been addressed by any previously published system. The framework employs a LangGraph-orchestrated pipeline (LangChain, Inc., 2024) of independent semantic and structural LLM agents combined with a deterministic scoring module, implementing a hybrid reasoning paradigm that leverages the complementary strengths of neural contextual flexibility and symbolic domain constraints.

The system was evaluated on a corpus of 99 expert-annotated hieroglyphic sequences under a nested cross-validation protocol: 5-fold outer cross-validation for generalisation estimation, with each outer training partition wrapped in an inner Dietterich 5 $\times$2 cross-validation that drives a Bayesian (Tree-structured Parzen Estimator) search over the parameters of the deterministic scoring module. The full multi-agent pipeline was executed on each held-out test fold. Every sequence in the corpus contributes to the final performance estimate as a held-out test case exactly once.

Across the five outer folds, the proposed method achieved a macro-averaged F1 of 81.94% $\pm$2.19% and a micro-averaged F1 of 87.50% $\pm$2.50%, with macro-precision of 87.20% $\pm$2.78% and macro-recall of 86.72% $\pm$5.07%. The low cross-fold variance indicates that the results do not depend on a fortunate train/test split and that the system's performance is structurally stable. The gap between inner-CV validation (90.4%) and held-out test (81.9%) macro-F1 is itself consistent across folds (std 3.0 pp), giving a quantified estimate of the cost of the LLM inference path relative to the rule-only path that the inner CV can directly optimise. Performance analysis revealed a complexity paradox consistent across all folds: the system performed best on information-rich sequences with three or more determinatives, where evidence triangulation from multiple independent sources produced robust convergent classifications, while sparse sequences with a single determinative or none, FP Null and FN Null cases, remained the dominant residual challenge.

The framework introduced several methodological innovations with implications beyond Egyptological NLP. The independence of the semantic and structural agents ensures that their agreement constitutes genuine evidential convergence analogous to inter-rater reliability, rather than shared bias. The five-tier sign categorisation system and context-dependent semantic thresholds anchor LLM-generated predictions to curated domain knowledge. The recently-added penalty mechanism enables conservative, organic growth of the knowledge base without inflating false positives.

Perhaps most significantly, the system departs from the standard single-output classification paradigm by generating ranked alternative interpretations with explicit promotion conditions. The promotion conditions, specifying the minimal evidence needed to prefer an alternative over the main interpretation, transform competing hypotheses into testable scholarly propositions, creating a natural feedback pathway for system improvement through domain expert interaction.

The system's full explainability, providing natural-language justifications, rejected-candidate reasoning, and ambiguity analysis for every classification decision, is essential for adoption in a field where computational tools must earn trust through

transparent, contestable reasoning. Rather than replacing expert judgement, the proposed method functions as a collaborative tool that accelerates the annotation of hieroglyphic corpora while preserving the scholar's interpretive authority.

Several future works emerge from the current evaluation. Enhanced contextual modelling through sequence-level embeddings could improve discrimination for dual-function signs such as D54 and A1. Knowledge refinement or caching mechanisms could reduce computational cost, enabling interactive use. Evaluation on diachronically diverse texts (Old Kingdom, Late Period, demotic) would assess robustness to orthographic variation. Finally, weighted evaluation metrics reflecting the differential Egyptological significance of determinative types would better capture the real-world impact of classification errors. The word boundary segmentation module, which performed reliably across all five folds, could be extended and evaluated independently as a general-purpose MdC text segmentation tool.

By establishing a strong baseline for automatic determinative detection and introducing a multi-agent architecture that balances precision with interpretive richness, the approach opens a new avenue in computational Egyptology and contributes a reusable framework for applying multi-agent LLM systems to other tasks involving the analysis of ancient writing systems.

Acknowledgements. The authors thank the Egyptologist collaborators for providing expert annotations and validating the knowledge base underlying this work.

Statements and Declarations

- **Funding.** This work forms part of the *Text Mining the Coffin Texts* project (CNS2023-144010, Research Consolidation Programme, funded by the Spanish Ministry for Science and Innovation).
- **Competing interests.** The authors have no competing interests to declare that are relevant to the content of this article.
- **Ethics approval and consent to participate.** Not applicable.
- **Consent for publication.** All authors have read and approved the manuscript and consent to its publication.
- **Data availability.** The expert-annotated benchmark of 99 hieroglyphic sequences produced and analysed in this study, together with the accompanying knowledge base and evaluation scripts, is publicly available in the Zenodo repository: https://doi.org/10.5281/zenodo.21776840. The development version is also maintained on GitHub at https://github.com/dcruzcavalieri/Multi_Agent_Framework_Egyptian_Hieroglyphic_Sequences.
- **Materials availability.** Not applicable.
- **Code availability.** All source code, configuration files, and experimental scripts required to reproduce the results are publicly available in the same repository as the data: https://doi.org/10.5281/zenodo.21776840.
- **Author contribution.** D.C.C. conceived the study, designed the methodology, developed the software, performed the validation and formal analysis, and wrote the main manuscript text. P.S.S.J. contributed to the software development and methodology. C.G.Z. and J.G. provided the expert Egyptological annotations and

validated the knowledge base, with J.G. also curating the data. C.G.Z. and S.E.P.C. contributed to the conceptualization. C.G.Z., J.G., and S.E.P.C. reviewed and edited the manuscript. All authors read and approved the final manuscript.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A Next-Generation Hyperparameter Optimization Framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631. https://doi.org/10.1145/3292500.3330701.

Allen, J.P., 2000. *Middle Egyptian: An Introduction to the Language and Culture of Hieroglyphs*. Cambridge University Press.

Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Sheratt, J., de Freitas, N., 2022. Restoring and Attributing Ancient Texts Using Deep Neural Networks. *Nature* 603(7900), 280–283.

Barucci, A., Cucci, C., Franci, M., Loschiavo, M., Argenti, F., 2021. A Deep Learning Approach to Ancient Egyptian Hieroglyphs Classification. *IEEE Access* 9, 123438–123447.

Barucci, A., Cucci, C., Franci, M., Loschiavo, M., Argenti, F., 2023. Ancient Egyptian Hieroglyphs Segmentation and Classification with Convolutional Neural Networks. *Algorithms* 16(2), 79.

Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B., 2011. Algorithms for Hyper-Parameter Optimization. In: *Advances in Neural Information Processing Systems*, pp. 2546–2554.

Buurman, J., Grimal, N., Hainsworth, M., Hallof, J., van der Plas, D., 1988. *Inventaire des signes hiéroglyphiques en vue de leur saisie informatique: Manuel de codage des textes hiéroglyphiques en vue de leur saisie sur ordinateur*. Institut de France.

Cawley, G.C. and Talbot, N.L.C., 2010. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research* 11, 2079–2107.

Champollion, J.F., 1836. *Grammaire égyptienne, ou Principes généraux de l'écriture sacrée égyptienne*. Firmin Didot Frères.

Chen, W., Su, Y., Zuo, J., Yang, C., Yuan, C., Qian, C., Chan, C.M., Qin, Y., Lu, Y., Xie, R., Liu, Z., Sun, M., Zhou, J., 2024. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors. In: *Proceedings of the International Conference on Learning Representations (ICLR 2024)*.

Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20(1), 37–46.

De Cao, M., Gracia Zamacona, C., Palazuelos, S., Grotenhuis, J., 2024. Deep Learning Meets Egyptology: A Hieroglyphic Transformer for Translating Ancient Egyptian. In: *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*. Association for Computational Linguistics, pp. 71–86.

Dietterich, T.G., 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation* 10(7), 1895–1923. https://doi.org/10.1162/089976698300017197.

Fischer, H.G., 1986. *L'écriture et l'art de l'Égypte ancienne*. Presses Universitaires de France.

Franken, M. and van Gemert, J.C., 2013. Automatic Egyptian Hieroglyph Recognition by Retrieving Images as Texts. In: *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, pp. 765–768.

Fuentes-Ferrer, J. et al., 2025. Recognition of Egyptian Hieroglyphic Texts through Focused Generic Segmentation and Cross-Validation Voting. *Applied Soft Computing* 170, 112674.

Fuentes-Jimenez, D. et al., 2025. The OCR-PT-CT Project: Semi-Automatic Recognition of Ancient Egyptian Hieroglyphs Based on Metrics Learning. arXiv preprint arXiv:2512.24197.

Gardiner, A.H., 1957. *Egyptian Grammar: Being an Introduction to the Study of Hieroglyphs*. Griffith Institute.

Goldwasser, O., 1995. *From Icon to Metaphor. Studies in the Semiotics of the Hieroglyphs*. University Press Fribourg Switzerland and Vandenhoeck & Ruprecht.

Goldwasser, O., 2002. *Prophets, Lovers and Giraffes: Wor(l)d Classification in Ancient Egypt*. Harrassowitz.

Goldwasser, O., 2006. On the New Definition of Classifier Languages and Scripts. *Lingua Aegyptia* 14, 473–484.

Goldwasser, O. and Grinevald, C., 2012. What Are Determinatives Good For?. In: Grossman, E., Polis, S., Winand, J. (Eds.), *Lexical Semantics in Ancient Egyptian*. Widmaier, pp. 17–53.

Goldwasser, O., Harel, H., Nikolaev, D., 2019. iClassifier: A Digital Platform for the Study of Graphemic Classifiers. In: *Ancient Egypt, New Technology*. Brill, pp. 131–156.

Goldwasser, O., 2023. From Determinatives to Classifiers: Categorisation in the Ancient Egyptian Writing System. In: Polis, S. (Eds.), *Guide to the Writing Systems of Ancient Egypt*. Institut Français d'Archéologie Orientale, pp. 192–199.

Goldwasser, O. and Handel, Z., 2024. Introduction: Graphemic Classifiers in Complex Script Systems. *Journal of Chinese Writing Systems* 8(1), 2–13.

Goldwasser, O. and Soler, S., 2024. Semantic Classifiers (Determinatives) and Categorization in the Ancient Egyptian Writing System: Rules, List of Classifiers, and Studies by iClassifier on the Story of Sinuhe. *Journal of Chinese Writing Systems* 8(1), 34–58.

Gracia Zamacona, C., 2001. Les déterminatifs de mouvement et leur valeur linguistique. *Göttinger Miszellen* 183, 27–45.

Gracia Zamacona, C., 2015. The Two Inner Directions of the Ancient Egyptian Script. *The Birmingham Egyptology Journal* 3, 9–23.

Gracia Zamacona, C., 2016. Verbes sans limite, verbes à limite: étude préliminaire d'après les données des Textes des Sarcophages. In: Collombert, P., Lefèvre, D., Polis, S., Winand, J. (Eds.), *Aere perennius. Mélanges égyptologiques en l'honneur de Pascal Vernus*. Peeters, pp. 303–326.

Gracia Zamacona, C., 2024. *Los Textos de los Ataúdes del Egipto antiguo: variabilidad, legitimación y diálogo*. Society of Biblical Literature.

Gracia Zamacona, C., 2025. Modulating Semograms: Some Procedures for Semantic Specification and Re-Categorization in the Pyramid Texts and Other Mortuary Texts. In: Cerbelló, J. and Orriols, M. (Eds.), *Signs, Language, and Culture. The Semograms of the Pyramid Texts between Iconicity and Referential Reality*. Brill, pp. 449–483.

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., Zhang, X., 2024. Large Language Model Based Multi-Agents: A Survey of Progress and Challenges. In: *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*.

Harel, H., Goldwasser, O., Nikolaev, D., 2024. iClassifier: A Digital Research Tool for Corpus-Based Classifier Networks in Complex Writing Systems. *Journal of Chinese Writing Systems* 8(1), 79–102.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S.K.S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., Schmidhuber, J., 2024. MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework. In: *Proceedings of the International Conference on Learning Representations (ICLR 2024)*.

Jakobson, R., 1956. Two Aspects of Languages and Two Types of Aphasic Disturbances. In: Jakobson, R. and Halle, M. (Eds.), *Fundamentals of Language*. Mouton, pp. 55–82.

LangChain, Inc., 2024. LangGraph: Multi-Agent Orchestration Framework. https://www.langchain.com/langgraph.

Lincke, E.S. and Kammerzell, F., 2012. Egyptian Classifiers at the Interface of Lexical-Semantics and Pragmatics. In: Grossman, E., Polis, S., Winand, J. (Eds.), *Lexical Semantics in Ancient Egyptian*. Widmaier, pp. 55–112.

McDonald, A., 2002a. Animal Metaphor in the Egyptian Determinative System. Three Case Studies. Ph.D. thesis. Oxford University.

McDonald, A., 2002b. An Evil Influence? Seth's Role as a Determinative, Particularly in Letters to the Dead. *Lingua Aegyptia* 10, 283–291.

McDonald, A., 2004. Review of Orly Goldwasser, Prophets, Lovers and Giraffes. Wor(l)d Classification in Ancient Egypt. *Lingua Aegyptia* 12, 235–244.

McDonald, A., 2007. A Metaphor for Troubled Times: The Evolution of the Seth Deity Determinative in the First Intermediate Period. *Zeitschrift für Ägyptische Sprache* 134, 26–39.

McDonald, A., 2014. Review of Die Prinzipien der Klassifizierung im Altägyptischen (GOF IV/38) by Eliese-Sophia Lincke. *Journal of Egyptian Archaeology* 100, 514–519.

Nederhof, M.J., 2013. The Manuel de Codage Encoding of Hieroglyphs Impedes Development of Corpora. In: Polis, S. and Winand, J. (Eds.), *Texts, Languages & Information Technology in Egyptology*. Presses Universitaires de Liège, pp. 103–110.

Nederhof, M.J., 2021. Formatting of Ancient Egyptian Hieroglyphic Text. In: Gracia Zamacona, C. and Ortiz-García, J. (Eds.), *Handbook of Digital Egyptology: Texts*. Editorial Universidad de Alcalá, pp. 17–35.

Nikolaev, D., Grotenhuis, J., Harel, H., Goldwasser, O., 2024. Classifier Identification in Ancient Egyptian as a Low-Resource Sequence-Labelling Task. In: *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*. Association for Computational Linguistics, pp. 42–47.

Sahala, A. and Lincke, E.S., 2024. Neural Lemmatization and POS-Tagging Models for Coptic, Demotic and Earlier Egyptian. In: *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*. Association for Computational Linguistics, pp. 87–97.

Shafer, G., 1976. *A Mathematical Theory of Evidence*. Princeton University Press.

Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I., de Freitas, N., 2023. Machine Learning for Ancient Languages: A Survey. *Computational Linguistics* 49(3), 703–747.

Thesaurus Linguae Aegyptiae, 2024. Thesaurus Linguae Aegyptiae. Berlin-Brandenburgische Akademie der Wissenschaften & Sächsische Akademie der Wissenschaften zu Leipzig. https://thesaurus-linguae-aegyptiae.de.

Varma, S. and Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7(91), 1–8. https://doi.org/10.1186/1471-2105-7-91.

Vernus, P., 2003. Idéogramme et phonogramme à l'épreuve de la figurativité: les intermittences de l'homophonie. In: Morra, L. and Bazzanella, C. (Eds.), *Philosophers and Hieroglyphs*. Rosenberg & Sellier, pp. 196–218.

Vernus, P., 2022. Script and Figurativity: Modelling the Relationship Between Image and Writing in Ancient Egypt. In: Morenz, L.D., Stauder, A., Büma, B. (Eds.), *Wege zur frühen Schrift: Niltal und Zweistromland*. EB-Verlag Dr. Brandt, pp. 335–421.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D., 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In: *Proceedings of the International Conference on Learning Representations (ICLR 2023)*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D., 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A.H., White, R.W., Burger, D., Wang, C., 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv preprint arXiv:2308.08155.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y., 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In: *Proceedings of the International Conference on Learning Representations (ICLR 2023)*.

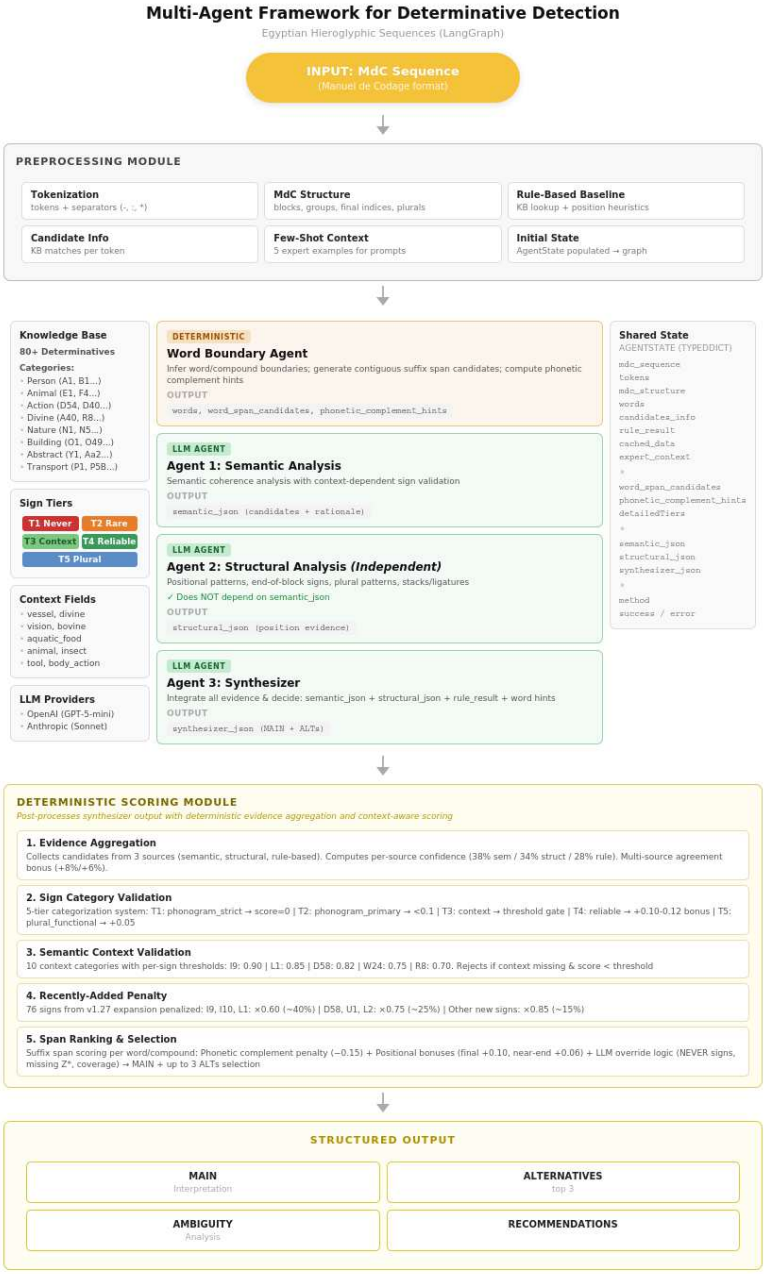


Fig. 1 Complete architecture of the proposed multi-agent framework for determinative detection. The pipeline comprises a preprocessing module, a deterministic Word Boundary Agent, two independent LLM agents (Semantic and Structural), a Synthesiser Agent, and a five-stage Deterministic Scoring Module, producing structured outputs with main interpretation, alternatives, ambiguity analysis, and recommendations.

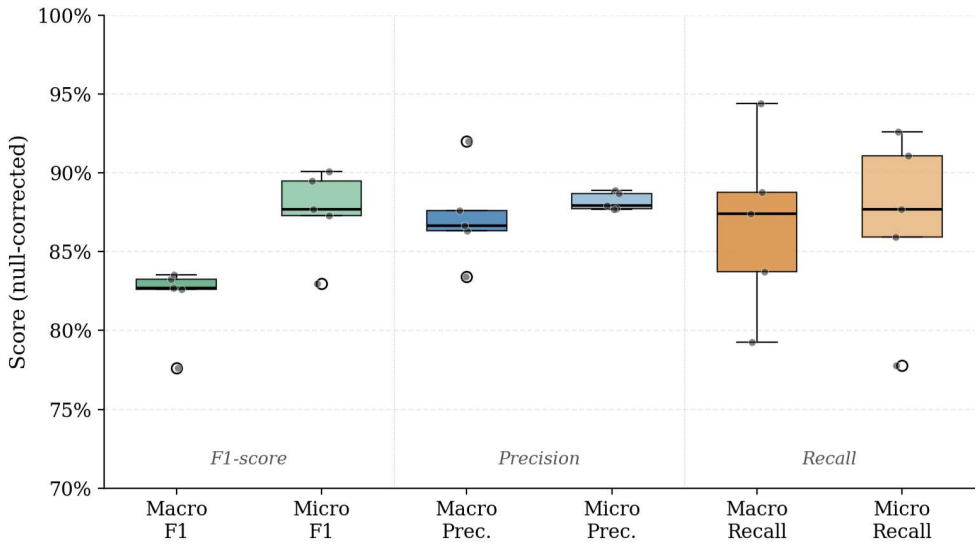


Fig. 2 Distribution of test performance across the five held-out outer folds for the proposed method. Each box summarises five fold-level values; individual fold scores are overlaid as points. The grouping shows F1, precision and recall side by side, with macro- and micro-averaged variants in each group. Micro variants pool all token decisions and are dominated by high-frequency signs; macro variants weight each sign class equally and therefore reflect performance on rare signs more strongly. The relative position of the macro and micro boxes within each metric is itself diagnostic where a larger spread between them indicates greater sensitivity of overall performance to the class distribution.

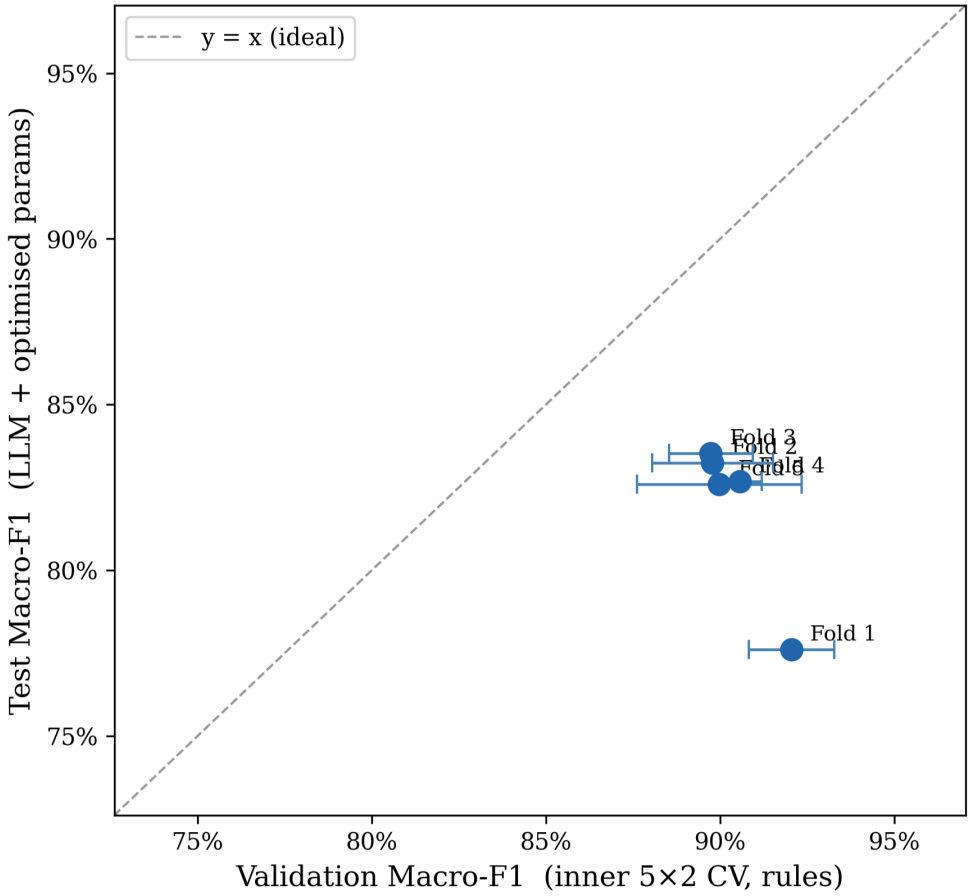


Fig. 3 Generalisation: inner-CV validation macro-F1 versus held-out test macro-F1 for each of the five outer folds. Error bars on x-axis show the standard deviation of validation F1 across the 5 $\times$2 inner repetitions. The dashed diagonal indicates $y = x$. All folds lie below the diagonal with a mean gap of -8.5 pp, indicating a consistent, and quantified, overestimation of test performance by the rule-only inner CV.

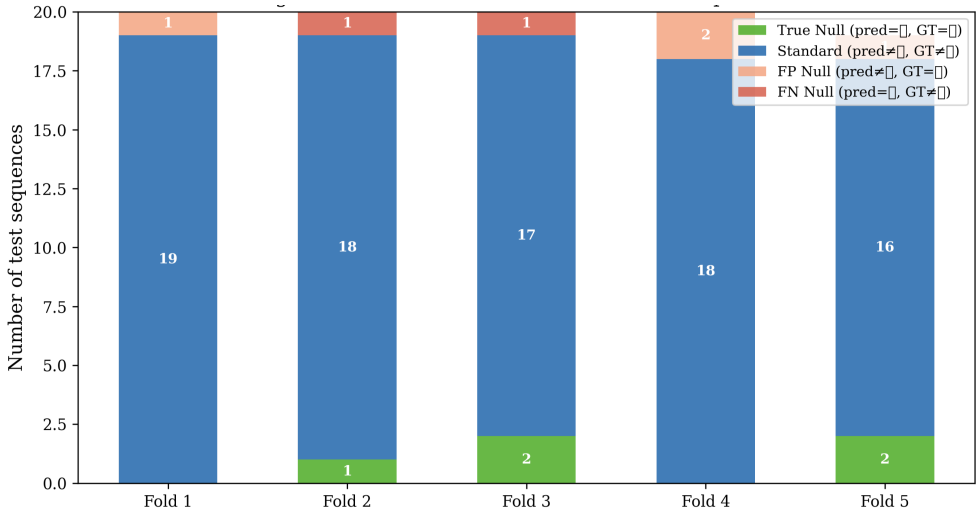


Fig. 4 Distribution of sequence-level outcomes across the five held-out test folds. STANDARD cases dominate every fold; TRUE NULL, FP NULL and FN NULL are reported separately to expose null-handling behaviour. The empty-sequence challenge, producing the correct abstention when no determinatives are present, remains the most variable component of the system's behaviour across folds.

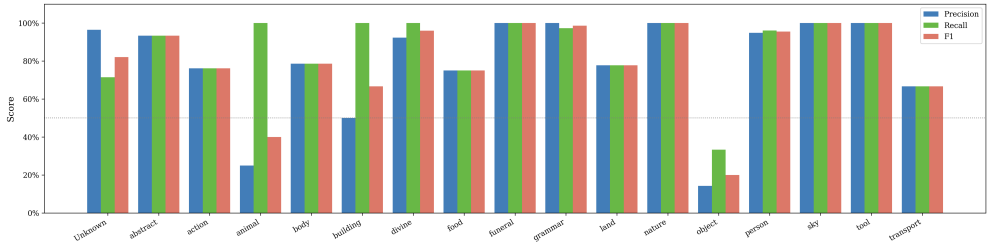


Fig. 5 Precision, recall and F1 per determinative category, pooled across all five held-out test folds. Categories such as FUNERAL, GRAMMAR, NATURE, PERSON and SKY reach $\geq 95\%$ F1, indicating that the system has internalised the characteristic contexts for these categories. The weakest categories are ANIMAL and OBJECT, where low precision suggests that LLM agents propose category-plausible candidates which are not present in the ground truth.

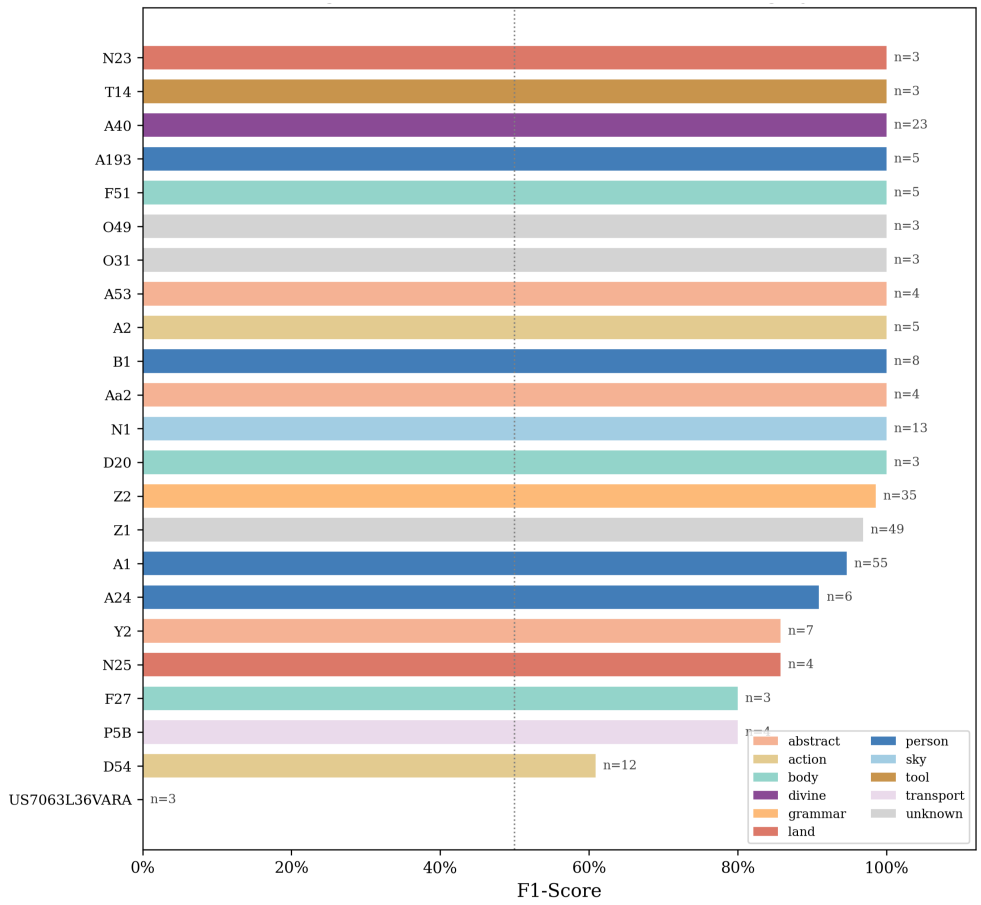


Fig. 6 Per-sign F1 for every sign with ≥ 3 ground-truth occurrences, pooled across all five held-out test folds. Colour denotes the sign's semantic category. The vast majority of common determinatives are detected at F1 ≥ 0.95; the two outliers are D54 (action category) and US7063L36VARA, the latter being a corpus-unique non-standard sign code absent from the knowledge base.