Supplementary Information

Mapping gender, agency, and objectification in Arabic poetry: an ontology-guided computational study

Anoud A. Alhamad[1], Wadee A. Nashir[2], Hassan Alshamrani[3,*], Badriyya B. Al-onazi[4], Reema G. Al-anazi[5], Wala F. Bin subait[4], Najla I. Al-shathry[4], Azizah M. Almutairi[4]

[1] Department of English Literature, College of Languages, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

[2] Digital Dhad Experts, Riyadh, Saudi Arabia

[3] Department of Arabic Language Studies as a Second Language, College of Languages and Sciences, King Saud University, Riyadh, Saudi Arabia

[4] Department of Language Preparation, Arabic Language Teaching Institute, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

[5] Department of Arabic Language and Literature, College of Humanities and Social Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

[*] Corresponding author email: rgalanizy@pnu.edu.sa

Contents of this Supplementary File:

- **Supplementary Note S1:** Epistemic Rationale and Verification Procedures (accompanied by Supplementary Fig. S1).
- **Supplementary Note S2:** Formal Semantic Representation and Interpretive Uncertainty (accompanied by Supplementary Fig. S2).
- **Supplementary Note S3:** Extraction Pipeline, Validation Algorithms, and Execution Parameters (accompanied by Supplementary Algorithm S1, Supplementary Fig. S3, and Supplementary Table S1).
- **Supplementary Note S4:** Extended Validation Profiles, Granular Agency Metrics, and Corpus Network Topology (accompanied by Supplementary Tables S2 and S3).
- **Supplementary Note S5:** Qualitative Error Casuistry, Propagation Topologies, and Cross-Domain Transferability (accompanied by Supplementary Fig. S4 and Supplementary Table S4).
- **Supplementary References**

Supplementary Note S1: Epistemic Rationale and Verification Procedures

This note expands the epistemic principles underlying the computational analysis of gendered agency and objectification in Classical Arabic poetry. Its central concern is the distinction between generating a plausible semantic interpretation and establishing that the interpretation is adequately grounded in the text.

The framework is intended to support corpus-scale literary analysis while keeping its representations open to inspection, contextual reconsideration, and scholarly critique. The overarching design philosophy and verification workflow connecting these methodological stages are illustrated in Supplementary Fig. S1.

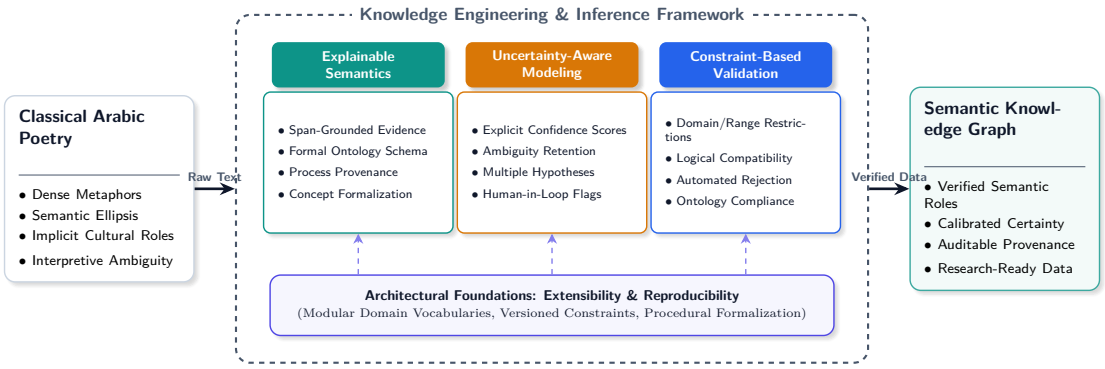


Figure S1: Supplementary Fig. S1: The methodological design philosophy of MARAYA. A conceptual overview linking the interpretive challenges of Classical Arabic poetry to evidence-grounded semantic representation, explicit uncertainty preservation, constraint-based validation, and human expert adjudication.

S1.1 Interpretive Traceability and Ontology Grounding In computational literary studies, aggregate patterns are more informative when they can be related to the textual material from which they were derived (Kuhn, 2019; Burrows, 2023). This is particularly important when the categories under examination, including agency, gaze, and objectification, are interpretive constructs rather than directly observable lexical properties. A computational label should not be treated as self-explanatory evidence of a literary phenomenon.

The framework therefore links extracted semantic claims to their supporting textual segments and represents them through the predefined concepts and relations of the study. This provides two complementary forms of grounding. Textual grounding allows an assertion to be examined against the relevant verse or passage, while ontology grounding makes the semantic category assigned to that assertion explicit. Formal ontologies can clarify the intended meaning and relationships of represented concepts (Guarino, 1995); they do not, however, determine whether every interpretation is persuasive in its literary or historical context. The distinction between the textual evidence, the category assigned to it, and the interpretation subsequently drawn from it must therefore remain available for critical examination.

S1.2 Representing Interpretive Uncertainty Classical Arabic poetry frequently employs metaphor, ellipsis, implicit reference, and rhetorical indirection. These features can leave the referent, gender attribution, or figurative meaning of an expression open to more than one plausible interpretation. Research on metaphor in Arabic likewise underscores the importance of contextual interpretation (Zibin et al., 2025). Consequently, uncertainty cannot be understood solely as an extraction failure: in some cases, it reflects a genuine limitation on what the available textual evidence can establish.

The framework treats uncertainty as relevant to the interpretation of extracted claims rather than assuming that every candidate must be resolved into an unconditional categorical assertion. Where the extraction procedure records confidence or ambiguity information, such information should be interpreted as an indication of the status of a computational hypothesis, not as a direct measure of literary truth. Referential, gender-related, and metaphorical ambiguity are useful distinctions for describing different

sources of interpretive difficulty; their use must remain consistent with the categories actually employed in the annotation and analysis. In particular, a confidence score does not by itself resolve competing readings, and a low-confidence interpretation is not necessarily an invalid one. The appropriate response depends on the textual evidence and the interpretive question under consideration.

S1.3 Structural Validation and Human Adjudication Probabilistic extraction and formal validation serve different purposes. The extraction stage proposes candidate semantic relations from textual context, whereas ontology-based constraints and rule-based checks assess whether represented assertions conform to the specified structural and semantic conditions. Where SHACL-based validation is used, it provides an executable means of checking graph data against declared constraints (Ke et al., 2024; Ferranti et al., 2024). These checks can identify violations of the representation requirements encoded in the validation rules; their scope is necessarily limited to the conditions those rules express.

Structural validity must therefore be distinguished from interpretive validity. An assertion may conform to the ontology while remaining debatable as a reading of a metaphor, an implicit reference, or a historically situated literary expression. Conversely, a structurally inadmissible assertion requires correction or review but does not, by that fact alone, settle the literary meaning of the underlying passage. Formal validation is thus a means of controlling the consistency of the representation, not a substitute for contextual interpretation.

Cases involving unresolved ambiguity, conflicting candidate interpretations, or concerns about the adequacy of textual support require human scholarly judgement where the study's review procedure provides for such adjudication. Review should consider the relevant textual evidence and the proposed relation, rather than treating a model-generated candidate as an established fact. Where multiple readings remain defensible, the limits of the available evidence should be acknowledged rather than resolved through unsupported certainty. This division of responsibilities keeps formal checks, probabilistic hypotheses, and human interpretation conceptually distinct.

S1.4 Transparency, Extensibility, and Reproducibility The interpretation of literary material depends on the corpus, the historical and generic scope of the inquiry, and the definitions used to operationalise its analytical categories. A semantic representation should therefore make its concepts and relations sufficiently explicit to permit scrutiny and, where justified, adaptation to other periods, genres, or research questions. Modular approaches to ontology design offer one means of accommodating such development without treating a representation as permanently fixed (Shimizu et al., 2023).

Reproducibility also requires attention to the decisions that shape computational outputs, not merely the reporting of aggregate evaluation measures (Pineau et al., 2021; Dodge et al., 2019). Relevant documentation includes the extraction procedure, semantic specifications, validation conditions, and review criteria insofar as they affect the resulting annotations and analyses. Claims about reproducibility or public availability should be limited to the materials and procedures actually documented and made accessible. These principles support the critical re-examination of computational findings while preserving the distinction between the consistency of a formal representation and the interpretive judgements on which literary analysis depends.

Supplementary Note S2: Formal Semantic Representation and Interpretive Uncertainty

This note specifies the event-centred representation used by MARAYA and clarifies the relationship between semantic modelling, uncertainty representation, and structural validation. The formalisation

makes the study's operational categories inspectable; it does not imply that literary interpretation can be reduced to a set of formal constraints or that structural conformance establishes interpretive truth.

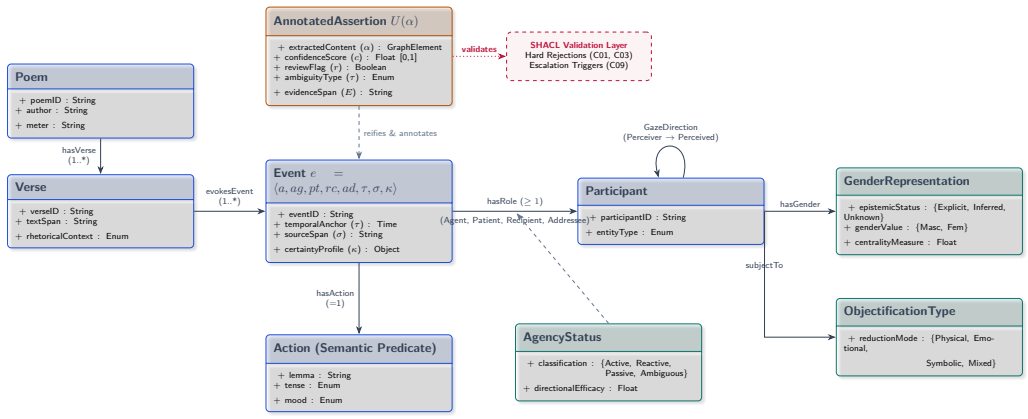


Figure S2: Supplementary Fig. S2: The core semantic ontology schema for classical Arabic poetry. This schema illustrates the operationalization of literary constructs (Gender, Agency, Objectification, and Gaze) within an event-centric representation.

S2.1. Event Representation and Structural Constraints MARAYA distinguishes a normalised action from an event representing its occurrence in a particular textual context. An action provides the semantic predicate, whereas an event associates that action with participants and contextual information relevant to its interpretation. This distinction is particularly important in Classical Arabic poetry, where metaphor, rhetorical address, syntactic compression, and zero-anaphora can affect how participants and their roles are recovered (Brown et al., 2017; Meghini et al., 2021).

Each event is represented by the following tuple:

$$e = \langle a, ag, pt, rc, ad, \tau, \sigma, \kappa \rangle \tag{S1}$$

where a denotes the action; ag the Agent; pt the Patient; rc the Recipient; ad the Addressee; τ the temporal anchor; σ the source span; and κ the confidence profile. The distinction between Recipient and Addressee is retained because the entity receiving a transferred object or outcome need not be the rhetorical target of an invocation or speech act. The tuple thus distinguishes participant roles from temporal, evidential, and confidence-related information.

The event representation is associated with structural requirements expressed in the original specification as:

$$\mathsf{Event} \sqsubseteq (= 1, \mathsf{hasAction}.\mathsf{Action}) \sqcap (\geq 1, \mathsf{hasParticipant}.\mathsf{Participant}) \tag{S2}$$

This expression specifies that an event has exactly one action and at least one participant of the indicated type. The original specification also describes a relationship between participants as:

$$\mathsf{Relationship} \sqsubseteq \mathsf{links}(\mathsf{Participant}_i, \mathsf{Participant}_j) \tag{S3}$$

The latter expression is retained as a schematic description of the relationship signature. It should not

be read as a standard Description Logic inclusion axiom without an explicit formal definition of links and its argument structure. These expressions summarise the intended semantic structure; their correspondence to the executable ontology and validation rules should be checked against the implementation.

S2.2. Operational Categories for Literary Relations Gender is represented through the categories:

$$\mathsf{Gender} \in \{\mathsf{Explicit}, \mathsf{Inferred}, \mathsf{Unknown}\} \tag{S4}$$

Explicit denotes an attribution directly supported by the text; Inferred denotes an attribution dependent on contextual or linguistic interpretation; and Unknown preserves cases where the available evidence does not support a sufficiently grounded assignment. Gender attribution is analytically separate from semantic-role assignment, so that an unresolved gender attribution does not automatically invalidate other textually supported relations.

Agency is operationalised as a participant's involvement in an event, with attention to causal contribution and meaningful control rather than grammatical subjecthood alone (Stuhler, 2024). The study distinguishes Active, Reactive, Passive, and Ambiguous cases. These categories respectively represent initiating action, responding to action, being affected without meaningful control, and cases in which the evidence permits competing readings.

Objectification is classified as Physical, Emotional, Symbolic, or Mixed. The categories distinguish corporeal, affective, symbolic, and co-occurring forms of representation. Gaze is represented through a directed relation between a Perceiver and a PerceivedEntity, with the categories Male-to-Female, Female-to-Male, Mutual, and Self-directed (Luo et al., 2024). These categories are operational definitions adopted for the analysis, rather than exhaustive accounts of the corresponding literary concepts.

S2.3. Explicit Representation of Interpretive Uncertainty Metaphor, ellipsis, and referential ambiguity may permit multiple plausible interpretations. Treating every candidate classification as certain would obscure this distinction. MARAYA therefore represents an uncertain assertion using the following tuple:

$$U(\alpha) = \langle \alpha, c, E, r, \tau \rangle \tag{S5}$$

where α denotes the candidate assertion, c its confidence score, E the supporting textual span, r an escalation flag for human adjudication, and τ the ambiguity category. The ambiguity categories specified in the original formulation are Referential, Gender, and Metaphorical.

This representation separates the proposed assertion from the evidence supporting it and from the indication that further adjudication may be required. The tuple records uncertainty-related information; it should not be interpreted as a probability model or a calibrated uncertainty measure beyond what the implemented scoring and routing procedures actually establish (Chen et al., 2019).

S2.4. Structural Validation and Its Interpretive Limits MARAYA separates the generation of candidate semantic assertions from their structural validation. OWL supplies the conceptual vocabulary and class and property relationships, while SHACL and SHACL-SPARQL are used to check candidate representations against specified constraints (Pareti and Konstantinidis, 2022; Ferranti et al., 2024). The validation described in the manuscript includes checks concerning evidence requirements, participant-role structure, and incompatible assignments.

These checks establish conformance to the constraints that have been encoded; they do not independently determine whether a metaphorical interpretation, an inferred participant, or an agency attribution

is the only defensible reading of a verse. A structurally admissible assertion may remain interpretively contestable, just as a structural violation identifies a problem in the representation without settling the literary meaning of the source passage.

The distinction between structural consistency and interpretive adequacy is therefore central to the use of the resulting corpus. Quantitative findings should be understood in light of the operational categories, the evidence supporting the annotations, and the documented limitations of computational interpretation. The formal representation is intended to make these decisions inspectable and open to critical re-examination, not to eliminate interpretive plurality.

Supplementary Note S3: Extraction Pipeline, Validation Algorithms, and Execution Parameters

This note documents the technical procedures supporting the controlled extraction workflow described in Section **??**. It specifies the sequence of extraction and validation operations, the structure of candidate semantic assertions, the formal routing procedure, and the computational environment used in the study. These procedures govern the admissibility and traceability of generated representations without establishing the literary correctness of individual interpretations.

S3.1 Extraction Pipeline and Execution Architecture The extraction procedure separates contextual inference from deterministic validation and expert adjudication. Its principal stages are textual anchoring, morphological analysis, ontology-guided candidate generation, structural and logical validation, uncertainty-based routing, and knowledge graph construction.

The source verse is retained in an archival representation, while a normalised representation supports orthographic matching and morphological processing. The verse or hemistich constitutes the principal extraction boundary. Morphological features, including clitics, person, number, and gender, provide contextual evidence for identifying explicit and omitted participants. Poet identity, historical period, and genre are retained as associated metadata.

The language model proposes contextual semantic relations within the vocabulary established by the ontology. These proposals are assembled into a candidate graph and subjected to formal checks before incorporation into the validated representation. Assertions that cannot be resolved through the specified validation procedures may be rejected or referred for expert adjudication.

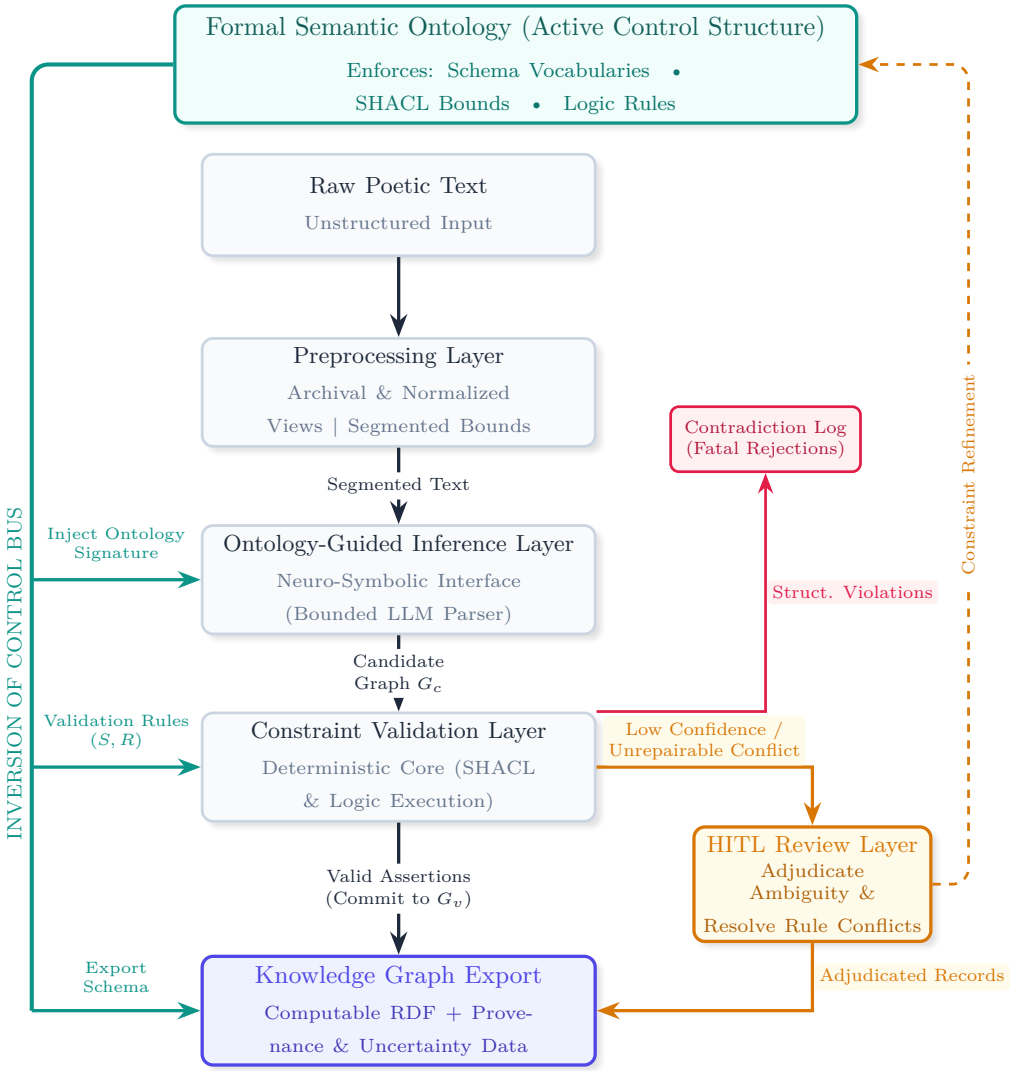
S3.2 Ontology Signature and Candidate Assertions Before contextual inference, an ontology signature is supplied through the extraction prompt. The signature specifies the admissible entity classes, semantic relations, controlled vocabularies, and relevant structural restrictions. Its purpose is to constrain the representational space within which the language model proposes interpretations.

The principal entity classes include Event, Participant, and Action. Event-centred relations connect participants to contextual actions and their associated semantic roles. The controlled literary-semantic vocabulary distinguishes gender attribution, agency, objectification, and gaze.

Gender attribution is qualified as Explicit, Inferred, or Unknown. Agency categories comprise Active, Reactive, Passive, and Ambiguous. Objectification categories comprise Physical, Emotional, Symbolic, and Mixed. Gaze direction is represented through Male-to-Female, Female-to-Male, Mutual, and Self-directed relations.

The candidate graph is represented as

$$G_c = \{(s, p, o, \alpha, e)\}, \tag{S6}$$



Supplementary Fig. S3. Extraction and validation architecture of MARAYA, showing textual anchoring, ontology-guided candidate generation, deterministic constraint checking, targeted expert adjudication, and knowledge graph construction.

where (s, p, o) denotes a subject–predicate–object assertion, $\alpha \in [0, 1]$ denotes its associated confidence value, and e identifies the supporting textual evidence span.

The confidence value and textual evidence remain associated with the candidate assertion during validation. Relations outside the controlled vocabulary are not admissible as domain relations. Conformity with the ontology signature nevertheless establishes neither textual entailment nor a uniquely defensible literary interpretation.

The source manuscript specifies the ontology-signature components and controlled categories but does not reproduce the complete literal extraction prompt. Accordingly, the description here documents the reported prompt structure without presenting a reconstructed prompt as the exact experimental input.

S3.3 Formal Constraint Validation Candidate assertions are evaluated against the ontology's structural and semantic requirements. SHACL is used to check required properties, domain and range compatibility, and cardinality constraints (Pareti and Konstantinidis, 2022). SHACL-SPARQL supports additional validation rules where dependencies cannot be expressed adequately through simpler structural constraints (Ferranti et al., 2024).

The validation procedure distinguishes structural admissibility from interpretive validity. For example, an agency assignment must be compatible with the event and participant-role structure defined by the ontology. Passing this check establishes conformity with the encoded conditions, not the correctness of the underlying literary reading.

Validation outcomes may include rejection, deterministic repair followed by renewed validation, referral for expert adjudication, or acceptance into the validated graph. The original routing procedure is reproduced in Supplementary Algorithm S1.

Algorithm 1 Supplementary Algorithm S1: Constraint-Based Validation and Routing

Input: Candidate graph G_c, Ontology O, SHACL constraints S, Logic rules R, Confidence threshold τ
Output: Validated graph G_v, Contradiction log L, HITL review queue H
Canonicalise identifiers in G_c to ontology IRIs
$V_1 \leftarrow \texttt{Validate_Shapes}(G_c, S)$
if Fatal structural violations detected in V_1 **then**
    $\texttt{Log}(V_1) \rightarrow L$; reject assertion
end if
$V_2 \leftarrow \texttt{Validate_Rules}(G_c, R)$
if Semantic contradictions detected in V_2 **then**
    $\texttt{Log}(V_2) \rightarrow L$
    if $\texttt{Repairable}(V_2)$ **then**
        Apply deterministic repair; re-validate
    else
        $\texttt{Route_to}(H)$
    end if
end if
for each surviving assertion $\alpha \in G_c$ **do**
    if $\texttt{confidence}(\alpha) < \tau$ **or** contains ReviewFlag **then**
        $\texttt{Route_to}(H)$
    else
        Commit $\alpha \rightarrow G_v$
    end if
end for

S3.4 Validation Outcomes and Expert Review The algorithm distinguishes fatal structural violations from semantic contradictions requiring further processing.

Fatal structural violations are recorded in the contradiction log and result in rejection of the affected assertion. Semantic contradictions are logged and assessed for deterministic repair. Where repair is possible, the affected representation is corrected according to the predefined rules and subjected to renewed validation. Where the contradiction cannot be resolved deterministically, the assertion is referred to the Human-in-the-Loop (HITL) review queue.

Assertions surviving structural and logical checks are also evaluated against the confidence threshold and any explicit review flags. Low-confidence or flagged assertions are referred for expert assessment rather than automatically incorporated into the validated graph.

The routing procedure is intended to concentrate scholarly attention on cases where formal processing is insufficient. Expert reviewers examine candidate interpretations alongside their textual evidence and may confirm, modify, reject, or retain unresolved readings. Their dispositions form part of the provenance

associated with the resulting semantic representation.

The confidence threshold τ is represented symbolically because its numerical value is not specified in the source section. The algorithm therefore documents the reported decision procedure without introducing an unverified numerical parameter.

S3.5 Knowledge Graph Representation and Provenance Following validation or targeted expert review, accepted assertions are incorporated into the knowledge graph. The representation connects event instances, participants, actions, semantic roles, and the literary-semantic categories defined by the ontology.

Each retained assertion is associated with its supporting textual evidence and relevant provenance information. The representation also preserves confidence status, validation outcome, and reviewer disposition where applicable. These records support subsequent inspection of the textual basis for corpus-level analyses.

The graph is intended to support structured comparison and evidence retrieval rather than to establish a single authoritative interpretation of each poetic construction. Ambiguous or contested interpretations therefore remain distinguishable from assertions admitted without unresolved review flags.

S3.6 Software Environment and Execution Parameters The computational environment was held fixed across the reported experimental conditions. The original execution configuration is reproduced in Supplementary Table S1.

Supplementary Table S1. Computational environment and model configuration.

Component	Configuration
Execution environment	Linux 6.6; Intel Xeon 2.20 GHz (2 vCPUs); 12.67 GB RAM
Software stack	Python 3.11; LangChain 0.2; RDFLib 7.0; pySHACL 0.25
Language model	Google Gemini 2.5 Flash API
Model parameters	Temperature $T = 0.1$; Top-$p = 1.0$; maximum output tokens 4096
Knowledge store	Apache Jena Fuseki; OWL 2 / RDF triple store

The reported model settings and software versions document the execution configuration used for the study. The fixed environment supports comparison across the evaluated conditions but does not, by itself, establish identical outputs across independent executions.

The extraction and validation procedures should therefore be interpreted together with the empirical consistency and expert-validation results reported in the main manuscript.

Supplementary Note S4: Extended Validation Profiles, Granular Agency Metrics, and Corpus Network Topology

This note reports supplementary class-level evaluation results, sequential changes in extraction performance, repeated-run stability, and descriptive properties of the corpus-scale knowledge graph. It complements the principal evaluation in Section **??** without treating formal consistency as a substitute for literary interpretation.

S4.1 Granular Agency Classification Breakdown Supplementary Table S2 presents precision, recall, and F1 for the four agency categories evaluated against the adjudicated expert reference. These results provide a more detailed view of the category-level agency F1 of 0.874 reported in the main text.

The Ambiguous category has the lowest F1 (0.761). This result is consistent with the interpretive difficulty of assigning agency where actions are indirect, metaphorical, or distributed among implied

Supplementary Table S2. Agency classification performance by granular class.

Agency class	Precision	Recall	F1
Active	0.914	0.902	0.908
Reactive	0.873	0.861	0.867
Passive	0.891	0.876	0.883
Ambiguous	0.781	0.742	0.761

participants. In such passages, a grammatical subject need not correspond straightforwardly to the participant understood as exercising literary agency.

The main error analysis reports that approximately 68% of semantic classification mismatches involved metaphorical participant attribution and indirect agency transfer, while 17% of overall errors concerned span-boundary mismatches. These findings help explain why agency classification cannot be reduced to surface grammatical roles. They do not, however, establish that the full 68% is attributable specifically to errors in the Ambiguous category.

The lower F1 for Ambiguous should therefore be interpreted as an observed evaluation result rather than as evidence that all ambiguous passages share a single error mechanism. Retaining this category is important because a forced assignment to Active, Reactive, or Passive could conceal uncertainty that remains relevant to scholarly interpretation.

S4.2 Sequential Performance Gains The ablation comparison distinguishes three automated configurations and a subsequent expert-curated reference.

Configuration A uses unconstrained language-model extraction. Configuration B adds ontology-guided semantic categories and representational restrictions. Configuration C additionally applies deterministic constraint validation. Configuration D represents the result following targeted Human-in-the-Loop (HITL) adjudication.

The observed F1 values are

$$F1_A = 0.767, \qquad F1_B = 0.833, \qquad F1_C = 0.884, \qquad F1_D = 0.898. \tag{S7}$$

The sequential differences, expressed in percentage points, are

$$\Delta F1_{A\to B} = (0.833 - 0.767) \times 100 = +6.6, \tag{S8}$$

$$\Delta F1_{B\to C} = (0.884 - 0.833) \times 100 = +5.1, \tag{S9}$$

$$\Delta F1_{C\to D} = (0.898 - 0.884) \times 100 = +1.4. \tag{S10}$$

The corresponding changes in Cohen's κ are

$$\Delta\kappa_{A\to B} = 0.781 - 0.692 = +0.089, \tag{S11}$$

$$\Delta\kappa_{B\to C} = 0.846 - 0.781 = +0.065, \tag{S12}$$

$$\Delta\kappa_{C\to D} = 0.864 - 0.846 = +0.018. \tag{S13}$$

The largest automated F1 gain occurs when ontology guidance is introduced. Adding constraint

validation produces a further gain while substantially reducing structural contradictions. These comparisons support the empirical contribution of the evaluated components under the reported experimental conditions.

The final transition requires a different interpretation. Configuration D incorporates expert adjudication and is not an independently operating automated model. Its improvement over Configuration C therefore measures the effect of expert-assisted curation on the resulting representation, not a gain in the predictive performance of the unaided extraction system.

These differences are sequential comparisons rather than independent causal estimates of each component under every possible configuration. In particular, the reported design does not isolate all potential interactions between ontology guidance, constraint validation, and expert review.

S4.3 Repeated-Run Stability The original evaluation reports 30 independent extraction runs on the hold-out set. Across these runs, the post-validation Constraint Satisfaction Rate (CSR) had a mean of 99.3% and a standard deviation of 0.24 percentage points. The Contradiction Rate (CR) had a mean of 0.5% and a standard deviation of 0.11 percentage points.

These statistics indicate relatively limited absolute variation in the two reported formal-consistency measures under repeated execution. They should not be interpreted as evidence that individual semantic assertions were identical across runs or that interpretive classifications were invariant.

The reported evaluation also indicates a coefficient of variation below 2.5% for the primary F1 measure. This provides an additional indication of aggregate performance stability, although it does not establish the stability of individual literary interpretations.

The distinction is relevant because a model may produce similar aggregate validation scores while differing in its treatment of particular metaphorical or ambiguous passages. Repeated-run consistency is consequently a useful technical property, but it does not replace expert inspection of contested assertions.

S4.4 Macro-Scale Knowledge Graph Topology After validation and curation, the framework was applied to the complete collection of 31,209 verses. Supplementary Table S3 reproduces the reported graph-scale characteristics.

Supplementary Table S3. Comprehensive knowledge graph statistics.

Topological metric	Value
Poems / verses processed	31,209
Structurally validated assertions	285,974
Participant entities	78,512
Event nodes	129,829
Typed semantic relations	184,380
Total graph nodes	208,341
Total graph edges	470,354
Average node degree	4.51
Largest connected component	86.2%

The graph contains 208,341 nodes and 470,354 edges, with an average reported node degree of 4.51. Its largest connected component accounts for 86.2% of nodes under the reported graph construction.

These values describe the scale and connectivity of the resulting semantic representation. They do not independently establish the literary significance of particular participants or relations. Graph connectivity may reflect the representational schema, shared entity types, repeated participant references, and the procedures used to construct links.

The gender-based centrality and clustering measures retained in the main text should therefore be interpreted as properties of the constructed semantic network rather than direct measures of historical social relations. Their literary relevance depends on inspecting the textual evidence associated with the represented participants and interactions.

S4.5 Interpretation and Reproducibility Boundaries The supplementary results support three distinct conclusions. First, the agency classification task exhibits uneven performance across its categories, with the greatest difficulty in the explicitly ambiguous class. Second, the successive methodological configurations produce measurable improvements in agreement with the expert reference, while the final expert-curated stage must remain analytically distinct from automated extraction. Third, repeated-run statistics and graph-scale measurements document technical properties of the representation without establishing the truth of individual literary readings.

These distinctions are necessary when using MARAYA for corpus-level interpretation. The knowledge graph supports systematic retrieval and comparison of evidence-linked relations, but the resulting numerical patterns remain conditional on the corpus, annotation definitions, extraction boundaries, and validation procedures.

The supplementary measurements should therefore be read alongside the qualitative interpretation and limitations in the main manuscript, particularly the restriction to single-verse semantic relations and the possibility of multiple defensible readings of figurative poetic language.

Supplementary Note S5: Qualitative Error Casuistry, Propagation Topologies, and Cross-Domain Transferability

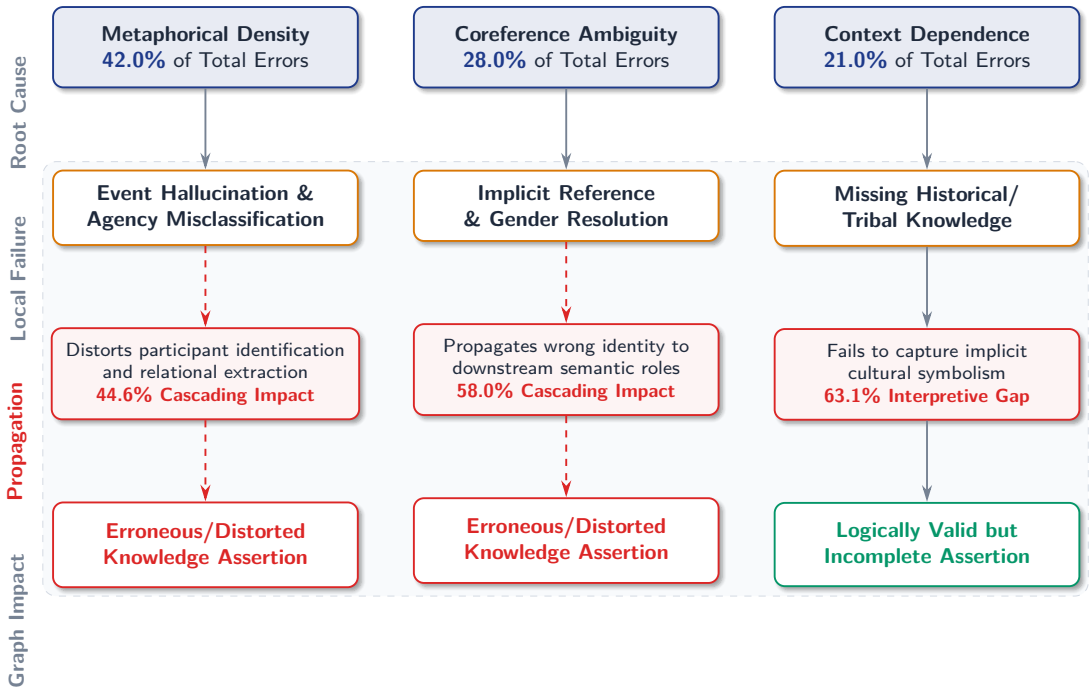
This note extends the discussion in Section **??** by examining how local extraction errors affect connected semantic assertions, clarifying the scope of the framework's empirical validation, and identifying methodological developments required to address the observed limitations. The analysis concerns the interpretive consequences of computational errors rather than treating every violation or disagreement as a purely technical defect.

S5.1 Error Cascade Topology and Failure Propagation The main-text taxonomy identifies 400 extraction errors classified by their primary cause. Because MARAYA represents literary relations through interconnected event, participant, and semantic-role assertions, an initial error may affect more than one element of the resulting graph.

Supplementary Fig. S4 illustrates this dependency structure. Errors in identifying an event or resolving a participant can alter subsequent agency, objectification, or gaze assignments. The figure represents possible pathways of error propagation rather than a claim that every initial error necessarily produces all downstream failures.

Metaphorical event propagation. Metaphorical Density accounted for 168 of the 400 analysed errors. Within this category, 104 cases involved Event Hallucination, in which figurative imagery was interpreted as a concrete event, and 64 involved Agency Misclassification, in which rhetorical positioning was mistaken for the assignment of agency.

These failures are analytically consequential because the event structure provides the basis for subsequent participant-role relations. For example, interpreting a personified image as an independently acting participant may produce a formally admissible event representation while changing the apparent distribution of agency in the verse.



Supplementary Fig. S4. Failure propagation in ontology-guided literary-semantic extraction. Local ambiguities in figurative event interpretation, participant identification, or contextual reference may affect downstream agency, objectification, and relational assertions.

In 75 of the 168 metaphor-related cases (44.6%), the initial misinterpretation propagated to at least one additional semantic component. This finding indicates that the consequences of metaphor-related errors extend beyond the original classification. It does not establish that every propagated case affected all relational categories or that the downstream effects were identical across cases.

Reference-resolution propagation. Coreference and Elliptical Ambiguity accounted for 112 errors, comprising 72 implicit or elliptical reference errors and 40 gender-resolution failures.

In Classical Arabic verse, a participant may be recoverable through grammatical agreement, discourse context, conventional address, or information supplied elsewhere in a poem. When the relevant antecedent is absent from the immediate textual span, a locally plausible resolution may nevertheless be inconsistent with the wider utterance.

An incorrect participant assignment may affect the interpretation of who initiates an action, who experiences its consequences, and the direction of an interpersonal gaze. In 65 of the 112 reference-related errors (58.0%), the initial error propagated to another semantic component.

These results demonstrate a dependency between participant identification and relational interpretation. They do not imply that all 65 cases simultaneously corrupted agency, objectification, and gaze.

Context-dependent interpretive divergence. The 84 Cultural and Historical Context Dependence errors comprised 38 historical allusions, 26 genealogical or entity-reference errors, and 20 cases involving cultural symbolism or conventional motifs.

These cases differ from straightforward structural failures. A candidate interpretation may contain admissible entities, valid relations, and correctly formed event structures while failing to recognise the significance of an allusion or the conventional function of a poetic image.

In 53 of these 84 cases (63.1%), the representation satisfied the applicable formal constraints but

differed from the expert-adjudicated interpretation. This finding demonstrates that the encoded constraints did not detect a substantial proportion of the observed context-dependent interpretive errors.

The result should be interpreted within its denominator: 63.1% refers to the 84 sampled context-dependent errors, not to all extracted assertions, all hold-out predictions, or the complete corpus. It therefore identifies a specific limitation of formal validation without measuring the overall prevalence of formally valid but interpretively incorrect assertions.

S5.2 Domain Applicability and Cross-Domain Transferability The ontology, extraction boundaries, and validation procedures were developed for the rhetorical and linguistic characteristics of Classical Arabic poetry. The framework's demonstrated performance must therefore be distinguished from its potential applicability to other textual traditions.

Supplementary Table S4 reproduces the applicability categories used in the original study.

Supplementary Table S4. Scope of empirically validated versus prospective applicability.

Target Textual Domain	Validation Status	Empirical Evidence and Adaptation Requirements
Classical Arabic Poetry	Empirically Validated	Direct experimental evaluation in this study.
Other Arabic Poetic Traditions	Prospective	Similar rhetorical features may support adaptation, but empirical revalidation is required.
Arabic Historical/Literary Prose	Requires Revalidation	Different discourse and event structures require empirical assessment and possible adaptation.
Non-Arabic Literary Corpora	Requires Revalidation	Application requires language-specific ontologies and empirical cross-lingual evaluation.

Classical Arabic poetry. The reported evaluation provides direct evidence for the framework within the corpus and annotation conditions examined. This validation does not imply uniform performance across every poet, period, rhetorical construction, or manuscript tradition.

Other Arabic poetic traditions. Traditions such as Andalusian *muwashshahat* and modern free verse may share figurative resources with the evaluated material, but their formal organisation, speaker structures, and discourse boundaries may differ. Adaptation would require reassessing the extraction unit, annotation categories, and treatment of extended or repeated references. Shared linguistic or rhetorical features alone cannot establish equivalent performance.

Arabic historical and literary prose. Prose narratives and *adab* literature may develop participant relations across longer textual sequences than the single-verse boundary adopted here. Their application would require renewed evaluation of event segmentation, discourse-level reference resolution, and the relevance of the existing literary-semantic categories.

Non-Arabic literary corpora. Cross-lingual transfer would require language-specific textual anchoring and morphological processing, together with assessment of whether the ontology's categories remain meaningful for the target tradition. Translation of category labels would not, by itself, establish conceptual equivalence or empirical validity.

These distinctions prevent the domain-specific evaluation from being presented as evidence of general literary or cross-lingual validity.

S5.3 Extended Methodological Horizons The error analysis suggests four complementary directions for future research. Each responds to an observed limitation but requires separate development and empirical evaluation.

1. Ontology evolution and interpretive category refinement.

The existing ontology provides a controlled representation of agency, objectification, gaze, and associated participant relations. Nevertheless, the observed difficulties with metaphorical agency and ambiguous reference suggest that some literary distinctions may not be adequately captured by a fixed category inventory.

Future ontology development could examine whether recurrent ambiguous cases warrant additional categories, explicit relations between competing readings, or more precise qualifications of agency and figurative attribution (Kotis et al., 2020).

Such refinement should not proceed by creating a new category for every difficult passage. Its scholarly value would depend on whether the revised distinctions are interpretable, consistently applicable, and supported by textual evidence. Any revised schema would also require renewed expert annotation and evaluation to establish whether greater representational detail improves rather than merely complicates analysis.

2. External historical and biographical knowledge grounding.

The cultural-context errors demonstrate that locally available linguistic evidence may be insufficient for interpreting historical allusions, genealogical references, and conventional motifs.

Future work could link candidate interpretations to curated historical commentaries, specialised lexica, biographical dictionaries, and classical *tabaqat* literature. Such resources could provide additional evidence for identifying named or implied referents and situating expressions within their relevant cultural traditions.

However, external knowledge would itself require provenance assessment. Historical sources may disagree, preserve competing attributions, or reflect interpretive traditions that should not be collapsed into a single authoritative reading. A knowledge-grounded system should therefore retain source attribution and represent unresolved disagreement rather than treating retrieved information as automatically decisive.

The proposed development would extend the evidentiary context available to reviewers; it would not eliminate the need for critical evaluation of historical sources.

3. Discourse-level participant and reference tracking.

The current single-verse boundary restricts the amount of discourse available for resolving omitted subjects, changing perspectives, and extended participant relations. This restriction supports manageable extraction units but limits the interpretation of narrative and thematic development across a complete *qasida*.

A future discourse-level model could track participants across successive verses, preserve changes in speaker and addressee, and distinguish locally expressed actions from relationships established through longer poetic sequences.

Such modelling would require an annotation protocol that explicitly addresses cross-verse reference and the continuity of literary participants. It would also need to accommodate uncertainty when a pronoun, image, or remembered figure admits more than one defensible antecedent.

Evaluation should distinguish improvements in reference resolution from improvements in the downstream interpretation of agency and objectification. Better identification of a participant would not necessarily resolve the literary significance of that participant's actions.

4. Explainable and interactive expert adjudication.

The divergence between formal consistency and expert interpretation demonstrates the importance of

interfaces that make computational decisions inspectable.

Future Human-in-the-Loop (HITL) tools could display the original verse, relevant morphological evidence, proposed event structures, supporting textual spans, model confidence information, and the specific constraints applied during validation. Reviewers could compare alternative interpretations and record the reasons for confirming, modifying, rejecting, or retaining an assertion as unresolved.

An effective interface should also distinguish a formal constraint violation from an interpretive disagreement. Presenting these outcomes as equivalent could encourage reviewers to mistake structural admissibility for literary correctness.

Interface evaluation would need to consider reviewer effort, consistency, and susceptibility to anchoring on machine-generated suggestions. The objective would be to support informed scholarly judgement rather than accelerate approval of automatically generated assertions.

S5.4 Implications for Corpus-Scale Interpretation The error analysis identifies three related but distinct sources of uncertainty: uncertainty about the event represented by figurative language, uncertainty about the participants associated with an event, and uncertainty about the cultural or historical interpretation of an otherwise admissible relation.

These uncertainties have different consequences for corpus-level findings. Event misinterpretation can introduce an unsupported relation; reference errors can attribute a relation to the wrong participant; and contextual errors can preserve a formally valid structure while assigning it a misleading literary meaning.

The reported propagation rates show that some local errors affect connected assertions, but they do not establish the magnitude of any resulting bias in the aggregate distributions of agency or objectification. Estimating that effect would require a separate analysis linking error categories to changes in corpus-level statistics.

Accordingly, the framework's formal checks, expert annotations, and provenance records should be treated as complementary safeguards. None is individually sufficient to guarantee interpretive correctness.

The principal methodological implication is that computational representations of literary relations are most useful when their evidentiary basis, uncertainty, and interpretive limits remain accessible to scholarly scrutiny.

References

Zibin, A., Binhaidara, N., Al-Shahwan, H., and Yousef, H. (2025). Metaphor interpretation in Jordanian Arabic, Emirati Arabic and Classical Arabic: Artificial intelligence vs. humans. *Humanities and Social Sciences Communications*, 12(1), 1–12. https://doi.org/10.1057/s41599-025-05282-0

Meghini, C., Bartalesi, V., and Metilli, D. (2021). Representing narratives in digital libraries: The narrative ontology. *Semantic Web*, 12(2), 241–264. https://doi.org/10.3233/SW-200421

Kuhn, J. (2019). Computational text analysis within the Humanities: How to combine working practices from the contributing fields? *Language Resources and Evaluation*, 53(4), 565–602. https://doi.org/10.1007/s10579-019-09459-3

Burrows, T. (2023). Reproducibility, verifiability, and computational historical research. *International Journal of Digital Humanities*, 5(2), 283–298. https://doi.org/10.1007/s42803-023-00068-9

Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. *International*

Journal of Human-Computer Studies, 43(5–6), 625–640. https://doi.org/10.1006/ijhc.1995.1066

Chen, X. et al. (2019). Embedding uncertain knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3363–3370. https://doi.org/10.1609/aaai.v33i01.33013363

Ferranti, N., De Souza, J. F., Ahmetaj, S., and Polleres, A. (2024). Formalizing and validating Wikidata's property constraints using SHACL and SPARQL. *Semantic Web*, 15(6), 2333–2380. https://doi.org/10.3233/SW-243611

Ke, J., Zacouris, Z., and Acosta, M. (2024). Efficient validation of SHACL shapes with reasoning. *Proceedings of the VLDB Endowment*, 17(11), 3589–3601. https://doi.org/10.14778/3681954.3682023

Shimizu, C., Hammar, K., and Hitzler, P. (2023). Modular ontology modeling. *Semantic Web*, 14(3), 459–489. https://doi.org/10.3233/SW-222886

Pineau, J. et al. (2021). Improving reproducibility in machine learning research (A report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164), 1–20.

Dodge, J. et al. (2019). Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194. https://doi.org/10.18653/v1/D19-1224

Brown, S. W., Bonial, C., Obrst, L., and Palmer, M. (2017). The rich event ontology. In *Proceedings of the Events and Stories in the News Workshop*, pages 87–97. https://doi.org/10.18653/v1/W17-2712

Stuhler, O. (2024). The gender agency gap in fiction writing (1850 to 2010). *Proceedings of the National Academy of Sciences*, 121(30). https://doi.org/10.1073/pnas.2319514121

Luo, K., Mao, Y., Zhang, B., and Hao, S. (2024). Reflecting the male gaze: Quantifying female objectification in 19th and 20th century novels. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13803–13812. https://doi.org/10.63317/5d5td7gmuc2b

Pareti, P. and Konstantinidis, G. (2022). A review of SHACL: From data validation to schema reasoning for RDF graphs. In M. Šimkus and I. Varzinczak, editors, *Reasoning Web. Declarative Artificial Intelligence*, pages 88–120. Springer. https://doi.org/10.1007/978-3-030-95481-9_6

Kotis, K. I., Vouros, G. A., and Spiliotopoulos, D. (2020). Ontology engineering methodologies for the evolution of living and reused ontologies: Status, trends, findings and recommendations. *The Knowledge Engineering Review*, 35, Art. e4. https://doi.org/10.1017/S0269888920000065