NOVEMBER 13, 2026 · V1.0.1

Formal Isomorphism Criteria for Cross-Disciplinary Terminology Crosswalks: A Three-Tier Verification Framework with Decidability and Sample-Size Analysis

Rowan Brad Quni-Gudzinas

QNFO-ULA

ABSTRACT

Cross-disciplinary knowledge transfer currently relies on intuition: researchers in different fields who work on the same structural object routinely fail to discover each other because their vocabularies do not overlap. A study would assume a partitionality index of 0.9721, with 453 of 466 technical terms occurring in exactly one domain and an empty shared core at the three-domain threshold. This paper addresses the gap between that lexical silence and the claimed structural unity of scientific concepts: no existing operational criterion distinguishes genuine cross-domain conceptual equivalence from superficial analogy. We formalize three connection types — structural isomorphism, functional equivalence, and logical interdependence — as a three-tier verification pipeline modeled on the L1/L2/L3 isomorphism hierarchy for attributed and dynamic networks. We derive the candidate search space for a six-domain vocabulary ($\binom{466}{2} = 108{,}345$ term pairs; $\binom{466}{4} \approx 1.94 \times 10^9$ relation quadruples), verify the projective invariance of a cross-ratio alignment score with a worked example ($\chi = 4/3$ preserved exactly under a Möbius transformation), derive the expert-validation sample size $n \geq 85$ from first principles, and show that a three-tier independence cascade reduces the false-positive rate from 10^{-2} per tier to 10^{-6} jointly. The framework converts crosswalk construction from an analogy-driven craft into a verification problem with explicit falsification criteria, and it is immediately applicable to terminology work in machine translation, terminography, and cross-disciplinary team collaboration.

PUBLISHER CLASSIFICATION High-quality scholarly research (QNFO publisher classification for open-access, versioned research with a public revision history and reader correction channel; this is not a claim of independent peer review.)

PEER-REVIEW STATUS Not independently peer-reviewed

CITE AS Quni-Gudzinas, R. B. (2026). Formal Isomorphism Criteria for Cross-Disciplinary Terminology Crosswalks: A Three-Tier Verification Framework with Decidability and Sample-Size Analysis. QNFO Preprint v1.0.1.

1. Introduction

A researcher in quantum foundations, a researcher in algebraic number theory, and a researcher in machine learning can work on the same structural object — a nested, hierarchical partition of a state space — and never discover each other [9]. This is not a rhetorical flourish but a measured condition: in a sample of six arXiv categories spanning quantum physics, number theory, machine learning, genomics, materials science, and economic theory, 97.2% of extracted technical vocabulary was domain-local, and not a single technical term was shared by three or more domains [9]. The people who face this problem are interdisciplinary researchers, terminologists, machine-translation providers, and the embedded cross-disciplinary software teams whose collaboration frictions have been documented in detail: in one activity-theoretic study of cross-disciplinary software development, expectations mapped to the *Community* component of Activity Theory appeared in 18 of 21 documented frictions, a node weight of 0.857 of all frictions in the study [1]. The cost is concrete — mistranslated domain terms, redundant research programs, collaboration friction, and terminology resources whose own bridge subsections "name connections without instantiating shared vocabulary" [9].

What blocks progress today is the absence of a principled criterion. Cross-domain concept alignment is currently done by analogy: a term in domain A "reminds" an expert of a term in domain B, and the mapping is defended rhetorically. Analogy has no error model. Without one, a terminology crosswalk — a validated mapping table between the term inventories of two or more domains — cannot be trusted, cannot be audited, and cannot be produced at scale by a computational pipeline. The decision that would change if this paper's result holds is the decision facing every terminologist, translation-system builder, and research-program designer: whether a proposed cross-domain term alignment is *verified* (backed by a checkable structural, functional, or logical certificate) or merely *suggested* (backed by surface similarity). A verified crosswalk changes machine-translation constraint injection, which today depends on manually curated term lists whose absence forces human translators to "resort to reference materials for terminology precision" [4]; it changes terminography, where generative tools now produce definitions that many users consider good enough but whose cross-resource consistency is unverified [8]; and it changes the onboarding and architecture decisions of cross-disciplinary teams, whose misaligned expectations are a documented friction source [1].

The core idea of this paper is that the three connection types an expert intuits — "these two concepts have the same structure," "these two concepts play the same functional role," and "these two concepts logically entail each other" — can be formalized as three independently checkable criteria, and that requiring *conjunction* rather than disjunction of the criteria separates genuine structural equivalence from superficial analogy. The idea has a natural formal home. A hierarchical framework for network comparison already distinguishes L1 (topological) isomorphism as "a perfect, adjacency-preserving bijective mapping between the vertex sets of two graphs," L2 (attributed) isomorphism, which additionally demands that "corresponding components must have identical properties," and L3 (dynamic) isomorphism as complete behavioral equivalence [10]. We adapt this hierarchy from networks to conceptual systems: concepts are vertices, their internal definitional structure and relations are edges and attributes, and a crosswalk is a mapping that must survive progressively stronger equivalence tests. Where the underlying hierarchy names the tiers by their network-theoretic origin (topological, attributed, dynamic), this paper names the transported criteria by their terminological function (structural isomorphism, functional equivalence, logical interdependence); the two namings denote the same three-tier conjunction, and we use the functional naming throughout.

This paper makes the following contributions:

- A formalization of the three connection types (structural isomorphism, functional equivalence, logical interdependence) as tiered criteria C_1, C_2, C_3 over attributed concept graphs, with a conjunctive decision rule and an independence-based false-positive cascade (Section 3, quantified in Section 4).
- Explicit derivations of the computational search space for a real six-domain vocabulary: $\binom{466}{2} = 108{,}345$ term pairs and $\binom{466}{4} = 1.94 \times 10^9$ relation quadruples, with independent verification of the $\binom{10^4}{4} \approx 4.164 \times 10^{14}$ quadruple budget reported in the cross-ratio literature (Section 4).
- A worked verification that the cross-ratio invariant is exactly preserved under Möbius transformation for a concrete embedding quadruple ($\chi = 4/3$ before and after), establishing that a projective alignment score can serve as a transformation-stable tier-1 statistic (Section 4).
- A first-principles derivation of the expert-validation sample size $n \geq 85$ aligned pairs, reproducing the requirement stated in the cross-ratio literature from stated assumptions about effect size, significance level, and power (Section 4).
- A quantitative ceiling on the lexical-overlap baseline: under the measured silo statistics, a surface-similarity pipeline can only ever propose alignments through the 13 bridge terms of the six-domain sample, bounding its reachable recall fraction at 0.0279 of the term inventory (Section 4).
- A falsification protocol and a proposed minimal empirical study that would test the framework's precision against expert-validated alignments, with the refutation criterion stated in advance (Sections 6 and 7).

Beyond terminology engineering, the framework bears on any setting where two formal or semi-formal vocabularies must be aligned under verification: specification formalization for verification-aware programming languages, where the correctness of a formal specification "cannot be defined mechanically" from intent alone and automated proxies for expert labels are actively sought [6]; and information retrieval model analysis, where the axiomatic structure of classical models is used as a diagnostic even when neural models violate it [7].

2. Background and Related Work

The vocabulary-wall measurement. The quantitative audit that motivates this work sampled the 40 most recent abstracts from each of six arXiv categories (quant-ph, math.NT, cs.LG, q-bio.GN, cond-mat.mtrl-sci, econ.TH), fetched 2026-08-24, and extracted 466 distinct technical bigrams [9]. Of these, 453 occurred in exactly one domain, 13 in two or more, and zero in three or more, yielding a partitionality index of 0.9721 and confirming the silo hypothesis H-SILO-1 (at least 90% of vocabulary domain-local) at 97.2% [9]. The same audit found that within a research program, semantic relatedness is lexically visible — same-program linked paper pairs share title vocabulary at mean Jaccard 0.115 against 0.028 for random pairs — while the knowledge-graph link infrastructure itself is program-local, with 8 same-program and 0 cross-program linked pairs among 40 classifiable pairs [9]. This asymmetry is the precise gap this paper targets: shared vocabulary tracks shared structure *within* a domain, but no mechanism currently certifies structural sharing *across* domains when the vocabulary is silent. The 13 bridge terms are dominated by method vocabulary (machine-learning, language-model, language-models, foundation-models, upper-bound — 5 of 13), which indicates that even the existing lexical bridges are methodological rather than conceptual [9].

The isomorphism hierarchy. The framework we adapt distinguishes three levels of network equivalence [10]. L1 isomorphism is defined as a perfect, adjacency-preserving bijective mapping between vertex sets — "the gold standard of absolute structural identity" — and its verification is the graph isomorphism problem, for which the chapter contrasts the theoretical difficulty of a universal efficient algorithm with the practical effectiveness of existing methods [10]. L2 isomorphism extends L1 to attributed graphs, formalized as a 4-tuple including attribute assignment functions and their attribute spaces, so that "corresponding components must have identical properties"; this "transforms the analysis from a comparison of abstract blueprints to a comparison of concrete, functional architectures" [10]. L3 isomorphism is complete behavioral or dynamic equivalence, the third and final tier of the hierarchy [10]. The same source frames the central tension this hierarchy resolves: similar topological patterns (scale-free, small-world) recur across domains, yet "a social network and a quantum system, even if they shared an identical blueprint, would behave in fundamentally different ways" [10]. This tension is exactly the analogy-versus-isomorphism distinction of the present research question: an L1 match alone is a blueprint analogy; only the conjunction of tiers certifies equivalence. Our contribution is to transport this hierarchy from networks to terminology.

Projective invariants as alignment statistics. The cross-ratio framework proposes an invariant $\chi(A, B; C, D)$ computed over word-embedding quadruples as a measure of translation fidelity and analogical structure, motivated by the observation that "existing similarity measures such as cosine distance ignore projective structure" [11]. Its reported results include a sentence-retention proportion $p_S = 0.511$, a token-retention proportion $p_T = 0.281$, a projected overall decoding speedup of approximately 510% under an independence assumption, a preserved $\chi = 1.25$ under the Möbius transformation $T(z) = \frac{2z+1}{z+3}$, an anharmonic orbit $\{1.25, 0.8, -0.25, -4, 5, 0.2\}$, a quadruple budget of 4.164×10^{14} candidates for $V = 10{,}000$, a sample-size requirement of $n \geq 85$ aligned pairs for a correlation test, and a processing-time estimate of approximately 3.5 hours for the cross-ratio pipeline on the weighted corpus [11]. Its own proposed falsification experiment computes the invariant over a parallel corpus and refutes the hypothesis if the Pearson correlation with sentence-level BLEU satisfies $|r| < 0.1$ [11]. We adopt the cross-ratio as a candidate tier-1 statistic precisely because it is invariant under the projective group, and we verify its invariance with our own worked example in Section 4.

Terminology principles and their computational turn. The operationalization of terminology for IT applications has revived the Wusterian approach, in which "the conceptual dimension once more prevails," demonstrated by the emergence of ontology in terminology; the Terminology Principles as defined in the Felber manual and the ISO standards remain at the core of traditional terminology while their computational form is re-examined [3]. This is the theoretical license for our tier-2 criterion: if concepts — not words — are the objects of terminology, then a crosswalk must align concepts, and concept identity is a structural property that can in principle be checked. The definitional side of the same tradition is under pressure from generative AI: the classical approach, embodied in ISO 704:2022, holds that the necessary and sufficient characteristics of a defined concept are universal and context-independent, yet these are "most often impossible to determine objectively" [8]. Generative tools can describe meaning activated "in a specific usage event" rather than a premeaning, and AI-assisted terminography including post-editing terminography — machine-produced, terminologist-corrected items — is the expected workflow [8]. The consequence for crosswalks is that tier-2 attribute matching cannot assume context-free concept definitions; it must score attribute agreement under stated contexts. The human-centered position is explicit in the terminology-work literature: generative AI creates automation opportunities but "large language models remain unreliable for terminological purposes," and efficiency pressures may encourage excessive automation on the perception that human expertise can be replaced [2]. Our framework is aligned with that position: the pipeline proposes, the tiers filter, and expert validation at $n \geq 85$ pairs (Section 4) remains the ground truth.

Terminology constraints in machine translation. The downstream consumer of a verified crosswalk is terminology-aware translation. The dominant obstacle is that neural machine translation systems fail to use "suitable domain-related words when translating specialized content not present in the training data" — the canonical example being "transformer," whose accurate translation depends on context or audience [4]. Prior work divides into soft constraints (inline injection of target terms, with lemmatization or grammatical-error-correction post-processing) and hard constraints (post-processing with desired terminologies, e.g., constrained beam search) [4]. A translate-then-refine system submitted to the WMT 2023 terminology task reported blind-test results including terminology recall and COMET QE scores, comparing recall with and without terminology constraints in random and real terminology scenarios [4]. A verified crosswalk would supply exactly the constraint lists such systems consume; the reported dependence of recall on constraint availability is the quantitative expression of the crosswalk's value.

Structural mapping as a transferable pattern. In neurodegeneration detection, multiscale structural mapping integrates gray-white matter contrasts with cortical thickness, and its extension MSSM+ adds sulcal and curvature features at the vertex level; MSSM+ yielded "approximately 3%p higher AUPRC and 1%p higher AUROC" than MSSM with lower across-fold standard deviations, and the associated supervertex models improved Alzheimer's-versus-cognitively-normal classification for each MR vendor [5]. The structural lesson transfers: adding features that carry genuine anatomical (i.e., domain-internal structural) information produced measurable, variance-reducing gains over the previous feature set. By analogy of method — not of content — our tier-2 attribute matching is the analogue of adding Sulc and Curv to CT and GWC: each added structurally grounded attribute should sharpen the separation between genuine and superficial alignments, and the MSSM+ result is the evidence that this design pattern pays in an adjacent domain [5].

Automated proxies for expert judgment. The problem of evaluating user-intent formalization for verification-aware languages (F*, Dafny, Verus) has no algorithmic solution, because intent is not mechanically checkable; the proposed remedy is to "symbolically 'test' the specifications" against validation tests, producing automated correctness and completeness metrics [6]. In implementation, a 400-line Python tool checked 64 of 153 specifications from the MBPP-DFY dataset, and the manual labels aligned with the tool's metrics for the large majority of examples: strong specifications scored completeness above 0.66 and weak ones below, "evidence that the automated metrics serve as a good proxy for

the user label" [6]. This is the closest existing template for our validation design: an automated tier score standing proxy for expert labels, with a threshold (there 0.66, here derived in Section 4) separating accepted from rejected alignments, and with the honest caveat that a minority of examples diverged.

Axiomatic diagnostics for neural matchers. White-box analysis of ColBERT on TREC-DL 2019 and 2020 (400 queries, BM25 candidates, $|\mathcal{S}_q| \leq 1000$) examined whether neural rankers respect classical IR axioms; a simplified ColBERT variant without query/document tokens or query augmentation obtained MRR@10 of 0.343 on MSMARCO dev versus 0.349, showing the simplifications were performance-negligible [7]. The findings: the IDF effect is not enforced, yet the model has a notion of term importance; exact matching remains important and is amplified by fine-tuning; and the axiomatic approach, "appropriate to analyze traditional IR models," remains "limited and somehow inadequate" for BERT-based models [7]. The relevance to this paper is methodological and cautionary: surface statistics of neural matchers (exact-match amplification, implicit term importance) are precisely the kind of analogy-grade evidence our tier criteria are designed to supersede, and the demonstrated inadequacy of classical axioms for neural models warns that our tier-1 structural tests must be validated, not assumed, for embedding-based representations.

Collaboration friction as the human cost. The activity-theoretic study of cross-disciplinary software development categorized collaboration models (professional, adjacent, and embedded forms), studied them with a mixed-method design including an exploratory interview at Microsoft, and constructed a network visualization of frictions in which vertex weight $w(v) = |\{c \in C \mid \exists e \in c \text{ s.t. } f(e) = v\}|$ counts frictions involving expectations mapped to Activity Theory component v; among 21 frictions, Community-related expectations occurred 18 times [1]. The study identifies friction hotspots to inform "future research, infrastructure design, and practical guidance for practitioners" [1]. A verified crosswalk is an infrastructure intervention at exactly this hotspot: a shared, auditable concept table reduces the expectation misalignment that the study documents as the mechanism of friction.

3. Methods

3.1 Object model

We model a domain's terminology as an attributed concept graph. Let a domain D be a tuple $D = (V_D, E_D, a_D, r_D)$ where V_D is the set of concept vertices (one per term), $E_D \subseteq V_D \times V_D$ is the set of typed conceptual relations (part-of, causes, measures, refines, contradicts), $a_D : V_D \to A$ assigns each concept a vector of attributes (definitional characteristics, typical contexts, measurement roles), and $r_D : E_D \to R$ assigns relation types. This mirrors the attributed-graph 4-tuple formalization of L2 isomorphism, which "explicitly includes attribute assignment functions and their associated attribute spaces" [10]. A cross-domain candidate alignment between domains D_1 and D_2 is a partial bijection $\phi : V_{D_1} \rightharpoonup V_{D_2}$.

3.2 The three tier criteria

Tier 1 — structural isomorphism (C_1). A candidate pair $(u, \phi(u))$ passes C_1 if the induced subgraph neighborhoods $N(u)$ and $N(\phi(u))$ are adjacency-preserving isomorphic under ϕ restricted to the neighborhoods, i.e., there exists a bijection $\psi : N(u) \to N(\phi(u))$ with $(x, y) \in E_{D_1} \Leftrightarrow (\psi(x), \psi(y)) \in E_{D_2}$ and $r_D(x, y) = r_D(\psi(x), \psi(y))$. This is the L1 criterion of adjacency-preserving bijection [10], applied locally to concept neighborhoods rather than globally to whole domains, because global concept-graph isomorphism across disciplines is not expected to hold and is not needed for a crosswalk.

Tier 2 — functional equivalence (C_2). A candidate passes C_2 if $a_{D_1}(u) = a_{D_2}(\phi(u))$ under a declared attribute dictionary $\delta : A_{D_1} \to A_{D_2}$, i.e., the attributes agree after dictionary translation. This is the L2 criterion that "corresponding components must have identical properties" [10], with the attribute dictionary playing the role of the attributed

graph's attribute spaces. Because classical definitional characteristics are "most often impossible to determine objectively" [8], we require attributes to be stated per declared context, and C_2 is evaluated per context; a pair passes only if it passes in all contexts declared for the crosswalk.

Tier 3 — logical interdependence (C_3). A candidate passes C_3 if the axioms in which u participates in D_1 and those in which $\phi(u)$ participates in D_2 are mutually interpretable: there is a translation τ of the axiom sets such that each translated axiom of D_1 is derivable from the axioms of D_2 and vice versa for the shared fragment. This is the transported form of L3 dynamic equivalence [10]: where L3 demands that processes on one network map to equivalent processes on the other, C_3 demands that inferential behavior (the "dynamics" of a conceptual system) is preserved. Complete behavioral equivalence is the strongest demand in the source hierarchy [10], and C_3 inherits that strength; in practice C_3 is evaluated on the shared axiom fragment, which is the portion of behavior a crosswalk can be held to.

Conjunctive decision rule. An alignment $(u, \phi(u))$ is *verified* if and only if it passes all three tiers:

$$\mathrm{Verified}(u, \phi(u)) \iff C_1(u, \phi(u)) \wedge C_2(u, \phi(u)) \wedge C_3(u, \phi(u)).$$

The conjunction, not any single tier, is what separates isomorphism from analogy: an L1-only match is exactly the "identical blueprint, different behavior" case the source hierarchy warns against [10].

3.3 The cross-ratio as a transformation-stable tier-1 statistic

For four embedded points A, B, C, D (e.g., word-embedding vectors projected to a common space), the cross-ratio is

$$\chi(A, B; C, D) = \frac{(C - A)(D - B)}{(C - B)(D - A)}.$$

The cross-ratio is invariant under the projective group, of which Möbius transformations $T(z) = \frac{az+b}{cz+d}$ are the one-dimensional complex representatives; the source framework reports preservation of $\chi = 1.25$ under $T(z) = \frac{2z+1}{z+3}$ and an anharmonic orbit $\{1.25, 0.8, -0.25, -4, 5, 0.2\}$ [11]. We use χ as the numeric core of C_1: two candidate pairs are structurally aligned when their neighborhood quadruples agree in cross-ratio, a condition stable under projective re-embedding of either domain.

3.4 Validation design

Ground truth for crosswalk evaluation is expert judgment; the pipeline's role is to propose and filter. This mirrors the validated proxy pattern of [6], where an automated completeness metric with threshold 0.66 served as "a good proxy for the user label" for the large majority of 64 checked specifications. The required expert sample size is derived in Section 4.4 rather than assumed.

3.5 Assumptions

1. The six-domain vocabulary statistics of [9] (466 terms, 453 domain-local, 13 bridges, partitionality 0.9721) are representative of the domains a crosswalk would target.
2. Tier false-positive rates are independent, so joint false-positive probability factorizes (Section 4.5); this is an idealization stated here and revisited in Section 6.
3. The minimum effect size worth detecting in validation is a Pearson correlation $r = 0.3$ between the pipeline score and expert labels, at significance level $\alpha = 0.05$ and power $1 - \beta = 0.8$; these are assumptions adopted to reproduce the $n \geq 85$ requirement of [11] from first principles.
4. Attribute dictionaries δ are declared before scoring, so that C_2 is decidable given the declared contexts.

4. Analysis

4.1 Search-space size for the six-domain vocabulary

The audit vocabulary contains $N_{\text{tot}} = 466$ distinct technical bigrams [9]. Unconstrained pairwise alignment search over the full inventory requires

$$\binom{N_{\text{tot}}}{2} = \frac{466 \times 465}{2} = \frac{216{,}690}{2} = 108{,}345$$

distinct term pairs to evaluate. This is the pairwise proposal space that any automated crosswalk pipeline over the six-domain sample must in principle traverse before tier filtering; the tier cascade exists precisely to cut this space down (Sections 4.5 and 4.6).

4.2 Relation-quadruple budget

Tier-1 scoring via the cross-ratio requires quadruples of embedded points (Section 3.3). Two quadruple budgets matter. For the audited vocabulary itself,

$$\binom{N_{\text{tot}}}{4} = \frac{466 \times 465 \times 464 \times 463}{24}.$$

Step by step: $466 \times 465 = 216{,}690$; $464 \times 463 = 214{,}832$; $216{,}690 \times 214{,}832 = 46{,}551{,}946{,}080$; dividing by 24 gives

$$\binom{466}{4} = 1{,}939{,}664{,}420 \approx 1.94 \times 10^{9}$$

relation quadruples. For the embedding vocabulary size $V = 10{,}000$ used in the cross-ratio literature [11], the budget is

$$\binom{10^4}{4} = \frac{10{,}000 \times 9{,}999 \times 9{,}998 \times 9{,}997}{24}.$$

Arithmetic: $9{,}999 \times 9{,}998 = 99{,}970{,}002$; $99{,}970{,}002 \times 9{,}997 = 999{,}400{,}109{,}994$; multiplying by $10{,}000$ gives $9{,}994{,}001{,}099{,}940{,}000$; dividing by 24:

$$\binom{10^4}{4} = 416{,}416{,}712{,}497{,}500 \approx 4.164 \times 10^{14},$$

which independently reproduces the quadruple budget of 4.164×10^{14} candidates reported for $V = 10{,}000$ [11]. The three-orders-of-magnitude gap between 1.94×10^{9} (the audited six-domain vocabulary) and 4.164×10^{14} (a full embedding vocabulary) is the quantitative reason the pipeline must filter tiers sequentially rather than score all quadruples at tier 1: tier-1 cross-ratio scoring at full embedding scale is the dominant cost, and every candidate eliminated by a cheaper pre-filter (e.g., the bridge-term ceiling of Section 4.6) removes $\binom{465}{3} \approx 1.67 \times 10^{7}$ quadruples downstream.

4.3 Worked verification of cross-ratio invariance

We verify the projective invariance that makes χ usable as a transformation-stable tier-1 statistic, with a concrete quadruple. Take $A = 0$, $B = 1$, $C = 2$, $D = 3$ on the real line. Then

$$\chi(A, B; C, D) = \frac{(C - A)(D - B)}{(C - B)(D - A)} = \frac{(2 - 0)(3 - 1)}{(2 - 1)(3 - 0)} = \frac{2 \times 2}{1 \times 3} = \frac{4}{3}.$$

Apply the Möbius transformation $T(z) = \frac{2z+1}{z+3}$, the same transformation under which the source framework reports preservation of $\chi = 1.25$ [11]:

$$T(0) = \frac{1}{3}, \quad T(1) = \frac{3}{4}, \quad T(2) = \frac{5}{5} = 1, \quad T(3) = \frac{7}{6}.$$

Recompute the cross-ratio on the images $A' = \frac{1}{3}$, $B' = \frac{3}{4}$, $C' = 1$, $D' = \frac{7}{6}$:

$$C' - A' = 1 - \frac{1}{3} = \frac{2}{3}, \qquad D' - B' = \frac{7}{6} - \frac{3}{4} = \frac{14 - 9}{12} = \frac{5}{12},$$

$$C' - B' = 1 - \frac{3}{4} = \frac{1}{4}, \qquad D' - A' = \frac{7}{6} - \frac{1}{3} = \frac{5}{6}.$$

Therefore

$$\chi(A', B'; C', D') = \frac{\frac{2}{3} \cdot \frac{5}{12}}{\frac{1}{4} \cdot \frac{5}{6}} = \frac{\frac{10}{36}}{\frac{5}{24}} = \frac{5}{18} \times \frac{24}{5} = \frac{24}{18} = \frac{4}{3}.$$

The cross-ratio is preserved exactly: $\chi = 4/3$ before and after the transformation, with zero numerical drift at exact arithmetic. Together with the reported preservation of $\chi = 1.25$ under the same map and the anharmonic orbit $\{1.25, 0.8, -0.25, -4, 5, 0.2\}$ [11], this establishes the invariance property on two independent quadruples and confirms that a tier-1 score built on χ is stable under projective re-embedding of either domain's concept vectors — the property a surface statistic such as cosine similarity lacks [11].

4.4 Expert-validation sample size

The validation design (Section 3.4) requires detecting a minimum Pearson correlation $r = 0.3$ between the pipeline score and expert labels at $\alpha = 0.05$ (two-sided) and power $1 - \beta = 0.8$ (assumptions 3, Section 3.5). Using the Fisher-z approximation for the correlation test, the required sample size is

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(\operatorname{arctanh} r)^2} + 3.$$

Inputs: $z_{0.975} = 1.96$, $z_{0.8} = 0.8416$, and

$$\operatorname{arctanh}(0.3) = \frac{1}{2} \ln \frac{1 + 0.3}{1 - 0.3} = \frac{1}{2} \ln \frac{1.3}{0.7} = \frac{1}{2} \ln(1.857142857) = \frac{0.619039}{2} = 0.309520.$$

Then $(\operatorname{arctanh} 0.3)^2 = 0.095802$, and

$$(z_{0.975} + z_{0.8})^2 = (1.96 + 0.8416)^2 = (2.8016)^2 = 7.84896,$$

$$n = \frac{7.84896}{0.095802} + 3 = 81.93 + 3 = 84.93.$$

Rounding up, the study requires

$$n \geq 85$$

expert-validated aligned pairs, reproducing from first principles the sample-size requirement stated in the cross-ratio literature [11]. The derivation makes explicit what the requirement was previously only asserted to be: it is the sample size at which a true pipeline–expert correlation of $r = 0.3$ is detected with 80% power, and it is the number of pairs the proposed falsification study of Section 7 must collect.

4.5 False-positive cascade under tier independence

Let each tier have false-positive rate $p_1 = p_2 = p_3 = 10^{-2}$ (assumption 2, Section 3.5: tier errors are independent, so probabilities factorize). Under the conjunctive decision rule, a superficial analogy is falsely verified only if it passes all three tiers simultaneously:

$$P_{\mathrm{FP}}^{\mathrm{joint}} = p_1\, p_2\, p_3 = (10^{-2})^3 = 10^{-6}.$$

The cascade therefore reduces the per-tier false-positive rate of 10^{-2} to a joint rate of 10^{-6}, a factor of 10^4. The independence assumption is an idealization: if tiers share failure modes (e.g., both tier 1 and tier 2 rely on the same embedding space), the joint rate is bounded below by $\max(p_1, p_2, p_3) = 10^{-2}$ and the 10^{-6} figure is an upper bound on the achievable reduction. We quantify the sensitivity in Section 6.

4.6 Ceiling on the lexical-overlap baseline

The audit found 13 bridge terms among 466, i.e., a bridge share of $13/466 = 0.027897 \approx 0.0279$ [9]. A surface-similarity pipeline can propose an alignment only where the two terms co-occur in some domain's text, i.e., only through bridge terms. Its reachable recall fraction over the term inventory is therefore bounded at

$$\rho_{\mathrm{lex}} = \frac{13}{466} = 0.0279.$$

Equivalently, in pair space: the number of pairs involving at least one bridge term is

$$\binom{466}{2} - \binom{453}{2} = 108{,}345 - \frac{453 \times 452}{2} = 108{,}345 - 102{,}378 = 5{,}967,$$

so a lexical pipeline can reach at most $5{,}967/108{,}345 = 0.0551$ of the pair space, and the remaining 94.9% of candidate pairs are invisible to it by construction. This is the quantitative ceiling on the analogy-only baseline: no amount of surface-similarity tuning can exceed recall 0.0279 of the inventory (or 0.0551 of pairs), because the vocabulary wall is a property of the corpora, not of the matcher [9]. Any crosswalk covering the other 97.2% of terms requires a non-lexical criterion — which is exactly what the three tiers supply.

4.7 Summary of derived quantities

Quantity	Value	Source
Pair search space $\binom{466}{2}$	108,345	Section 4.1
Quadruple budget $\binom{466}{4}$	1.94×10^9	Section 4.2
Quadruple budget $\binom{10^4}{4}$	4.164×10^{14}	Section 4.2, matches [11]
Cross-ratio invariance (worked example)	$\chi = 4/3$ preserved exactly	Section 4.3
Validation sample size	$n \geq 85$	Section 4.4, matches [11]
Joint false-positive rate	10^{-6} (upper bound)	Section 4.5
Lexical-baseline recall ceiling	0.0279 (inventory), 0.0551 (pairs)	Section 4.6

5. Results

The paper derives five results, each with fully shown arithmetic in Section 4.

R1 — Search-space sizes. The six-domain audited vocabulary of 466 terms induces $\binom{466}{2} = 108{,}345$ candidate term pairs and $\binom{466}{4} = 1{,}939{,}664{,}420 \approx 1.94 \times 10^9$ relation quadruples (Section 4.1–4.2). At embedding scale $V = 10{,}000$, the quadruple budget is $\binom{10^4}{4} = 416{,}416{,}712{,}497{,}500 \approx 4.164 \times 10^{14}$, independently reproducing the

figure reported in the cross-ratio literature [11].

R2 — Exact projective invariance. For the quadruple $(0, 1, 2, 3)$ and the Möbius transformation $T(z) = \frac{2z+1}{z+3}$, the cross-ratio equals $4/3$ before transformation and $4/3$ after, in exact arithmetic (Section 4.3). This verifies on a second independent quadruple the invariance property the cross-ratio framework reports for $\chi = 1.25$ [11], establishing χ as a transformation-stable tier-1 statistic.

R3 — Validation sample size. Detecting a pipeline–expert correlation of $r = 0.3$ at $\alpha = 0.05$, power 0.8, via the Fisher-z method requires $n = 84.93$, hence $n \geq 85$ expert-validated aligned pairs (Section 4.4), reproducing the requirement stated in [11] from stated assumptions.

R4 — False-positive cascade. Under tier independence, three conjunctive tiers each with false-positive rate 10^{-2} yield a joint false-positive rate of $(10^{-2})^3 = 10^{-6}$, a 10^4-fold reduction (Section 4.5). This is an upper bound on the achievable reduction; correlated tier failures bound the joint rate below by 10^{-2}.

R5 — Lexical-baseline ceiling. A surface-similarity pipeline restricted to the 13 bridge terms of the six-domain sample reaches at most a 0.0279 fraction of the term inventory and 0.0551 of the pair space ($5{,}967$ of $108{,}345$ pairs); the remaining 94.9% of pairs are unreachable by any lexical method (Section 4.6). This quantifies the gap the tier framework must close and bounds the value of any analogy-only baseline.

6. Discussion

What the results do and do not establish. The paper establishes a verification framework and derives its quantitative parameters; it does not establish that the framework achieves high precision or recall on real cross-domain corpora. R1–R5 are structural and statistical properties of the proposed pipeline, not empirical performance measurements. The single most important open empirical question is whether genuine cross-domain concept equivalences exist in sufficient number to make crosswalks worth building — the vocabulary audit found an empty shared core at the three-domain threshold [9], and if deep equivalences are rare, the pipeline will be precise but nearly empty.

Falsification. The main result — that conjunctive tier verification separates genuine equivalence from analogy — would be falsified by the following observation: an expert-validated study of $n \geq 85$ pairs (Section 4.4) in which the pipeline's verified alignments show no correlation with expert labels, operationally $|r| < 0.1$ between the pipeline score and expert judgment, the same refutation threshold the cross-ratio framework sets for its own hypothesis [11]. A secondary falsification path is internal: if tier scores are found to be strongly correlated across tiers (measured, e.g., by inter-tier phi coefficients near 1 on a labeled sample), the independence assumption of Section 4.5 fails, the joint false-positive rate rises from 10^{-6} toward 10^{-2}, and the cascade's central benefit disappears.

Limitations. The result rests on the following assumptions, with sensitivity noted for each:

- *Vocabulary representativeness.* The 466-term, 453-local, 13-bridge statistics come from 40 recent abstracts per domain in six arXiv categories [9]. If other domain pairs have higher bridge shares, the lexical ceiling of R5 rises and the framework's marginal value over lexical baselines falls proportionally; the ceiling is exactly the bridge share, so it is linearly sensitive to that statistic.
- *Tier independence.* The 10^{-6} joint false-positive rate factorizes only under independence (assumption 2). With pairwise tier correlation ρ, the joint rate interpolates between 10^{-6} ($\rho = 0$) and 10^{-2} ($\rho = 1$); the headline 10^4-fold reduction is therefore an upper bound, and the result degrades gracefully, not catastrophically, under moderate correlation.
- *Effect-size assumption.* The $n \geq 85$ figure assumes $r = 0.3$ is the minimum effect worth detecting. If the true pipeline–expert correlation is smaller, the required sample grows as $1/(\operatorname{arctanh} r)^2$: at $r = 0.2$, $\operatorname{arctanh}(0.2) = 0.2027$, giving $n \approx 7.849/0.0411 + 3 \approx 194$; the validation cost roughly doubles.

- *Context-relative attributes.* Tier 2 requires attribute agreement per declared context because classical definitional characteristics are "most often impossible to determine objectively" [8]; if the declared contexts are incomplete, C_2 can pass spuriously. The framework's honesty depends on the context declaration being auditable.
- *Embedding stability.* Tier 1 via χ presumes concept embeddings are stable enough that neighborhood quadruples are well-defined; the white-box analysis of ColBERT shows that classical axioms can be "limited and somehow inadequate" for BERT-based matchers even when surface behaviors look axiomatic [7], so tier-1 tests must themselves be validated on the embedding model in use, not assumed.
- *Proxy validity.* Expert labels remain ground truth, and the automated-proxy template on which the validation design is modeled found alignment "for the large majority" but not all of 64 checked specifications [6]; a minority divergence rate is expected here too and is not modeled in R4.

Failure modes. Three are foreseeable. First, false negatives: a genuine equivalence failing tier 2 because the two domains declare incompatible attribute dictionaries; the conjunctive rule is asymmetric, trading recall for precision by design. Second, dictionary gaming: whoever declares the attribute dictionary δ controls tier 2's difficulty, so crosswalk audits must publish δ alongside the crosswalk. Third, method-vocabulary contamination: the audit's bridge terms are dominated by method vocabulary (5 of 13) [9], so a lexical pre-filter tuned to bridges would over-propose methodological rather than conceptual alignments; the tier criteria must therefore be applied after, not instead of, a conceptual filter.

Relation to adjacent evidence. The design pattern of adding structurally grounded attributes to sharpen separation — validated in neurodegeneration detection, where adding sulcal and curvature features to cortical thickness and gray-white matter contrasts yielded approximately 3 percentage points higher AUPRC and 1 point higher AUROC with lower across-fold variance [5] — supports the expectation that each additional tier (a structurally grounded attribute of the alignment) improves separation, but the transfer is by analogy of method and is not itself evidence.

7. Practical Implications and Engineering Pathways

(a) Users and changed decisions. Three user groups would use a verified crosswalk. First, terminologists and terminographic resource builders: the framework gives them an auditable acceptance criterion for cross-resource concept mappings, replacing subsections that "name connections without instantiating shared vocabulary" [9] with tables whose entries carry tier certificates; the decision that changes is which candidate alignments enter a published resource. Second, machine-translation system builders: terminology-aware translation systems consume constraint lists, dividing prior work into soft constraints (inline injection with lemmatization or grammatical-error-correction post-processing) and hard constraints (post-processing such as constrained beam search) [4]; a verified crosswalk changes the provenance of those lists from manual curation to certified mapping, and the WMT 2023 terminology-task system that compared recall with and without constraints in random and real terminology scenarios [4] is the evaluation harness such lists would be dropped into. Third, cross-disciplinary software teams: the activity-theoretic study found expectations mapped to the Community component in 18 of 21 documented frictions [1]; a shared, auditable concept table is an infrastructure intervention at exactly that hotspot, changing the onboarding and architecture decisions such teams make.

(b) Readiness. This work is purely analytical and has not been peer reviewed; against the NASA Technology Readiness Level definitions it sits at TRL 1–2: the basic principles (the three-tier hierarchy transported from network isomorphism [10], the projective invariance of the cross-ratio [11]) are identified and the quantitative parameters are derived on paper, but no pipeline has been implemented, no corpus scored, and no expert validation run. Analytical work of this kind cannot exceed TRL 3, and no manufacturing or deployment readiness is claimed.

(c) Smallest next test. The minimal de-risking study, conditional in form: a study would select a publicly available parallel corpus, compute the cross-ratio score for each sentence pair, compute sentence-level BLEU, and apply the falsification criterion that a Pearson correlation $|r| < 0.1$ refutes the hypothesis that χ predicts translation adequacy [11]

— this tests tier 1's core statistic at a cost of one corpus and one correlation test. The terminology-level analogue: a study would recruit terminologists to validate $n \geq 85$ candidate alignments (Section 4.4) drawn from two of the six audited domains, would measure the correlation between the pipeline's verified/unverified decision and expert labels, and would compute inter-tier correlations to test the independence assumption; an outcome with $|r| < 0.1$ against expert labels, or inter-tier phi coefficients near 1, refutes the framework's central claims.

(d) Binding constraints. The first binding constraint is compute at tier 1: the quadruple budget of 4.164×10^{14} at embedding scale $V = 10{,}000$ (Section 4.2) makes exhaustive cross-ratio scoring infeasible, so candidate pre-filtering is mandatory; the processing-time estimate of approximately 3.5 hours for the cross-ratio pipeline on the weighted corpus [11] indicates the cost is manageable only after the sentence-level reduction to $p_S = 0.511$ of sentences and $p_T = 0.281$ of tokens [11]. The second binding constraint is expert bandwidth: validation requires $n \geq 85$ expert-judged pairs per evaluation (Section 4.4), and expert time, not compute, sets the pace of the validation loop. The third is data: the framework needs parallel multi-domain corpora with declared attribute dictionaries, and the audit's finding that the knowledge-graph link infrastructure is itself program-local (8 same-program, 0 cross-program linked pairs among 40 classifiable pairs) [9] indicates that the cross-domain link data a crosswalk pipeline would consume does not yet exist at scale. Regulation and cost drivers are open questions this paper does not quantify.

(e) Order-of-magnitude projections. Labeled as projections with stated assumptions. If tier-1 scoring costs c_1 per quadruple and the pre-filter cuts candidates to the bridge-reachable fraction 0.0551 of pairs (Section 4.6), the tier-1 workload falls from $\sim 4.164 \times 10^{14}$ to $\sim 2.3 \times 10^{13}$ quadruples at $V = 10{,}000$ — still large, which is why the six-domain vocabulary budget of 1.94×10^{9} (Section 4.2) is the realistic first target. These projections assume the per-quadruple cost is uniform and the pre-filter is independent of tier-1 errors; both are testable in the study of (c).

(f) Risks and misuse paths. A verified crosswalk inherits the authority of its verification: if the attribute dictionary δ is declared to make tier 2 trivially passable, the "verified" label launders analogy-grade mappings as certified ones. The mitigation is procedural — publish δ, the declared contexts, and the tier certificates with every crosswalk — and the human-centered position of the terminology literature applies: large language models remain unreliable for terminological purposes, and efficiency pressures may encourage excessive automation on the perception that human expertise can be replaced [2]; the pipeline proposes, the tiers filter, and expert validation remains the ground truth.

8. Conclusion

Cross-disciplinary terminology mapping has lacked an error model: analogy proposes alignments that nothing can check. This paper formalizes the three connection types an expert intuits — structural isomorphism, functional equivalence, logical interdependence — as a conjunctive three-tier verification pipeline transported from the L1/L2/L3 isomorphism hierarchy for attributed and dynamic networks, and derives its operating parameters in full: a pair search space of $108{,}345$ and a quadruple budget of 1.94×10^{9} for the audited six-domain vocabulary, exact preservation of the cross-ratio $\chi = 4/3$ under Möbius transformation as the tier-1 statistic's stability certificate, an expert-validation requirement of $n \geq 85$ aligned pairs derived from Fisher-z first principles, a joint false-positive rate of 10^{-6} under tier independence, and a hard 0.0279 recall ceiling on any lexical-overlap baseline. The framework converts crosswalk construction from analogy-driven craft into a verification problem with explicit falsification criteria, and the proposed minimal studies — a parallel-corpus correlation test for tier 1 and an $n \geq 85$ expert-validation study for the full pipeline — would confirm or refute its central claim with pre-stated thresholds.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.55.12). Language models were tools in that process; they are not authors. The named author is accountable for the content.

- Models: drafting by openai/gpt-oss-120b, zai-org/glm-5.3-flash; reconciliation, completion, review, revision and verification by zai-org/glm-5.3-flash, openai/gpt-oss-120b, moonshotai/kimi-k2.6, with the QNFO model router as a fallback.
- Sources: every reference is drawn from a bibliography retrieved from arXiv and the QNFO corpus before drafting; every cited work's full text was retrieved (arXiv HTML, or the QNFO paper's body) and the writers were given excerpts of its own sections; statements about a cited work are restricted to its retrieved title, abstract and those excerpts.
- Checks passed before publication: required sections present and finished, cross-references resolve, reference list rendered from the retrieved bibliography, no assistant residue or placeholders, adversarial model review with its blocking findings resolved, and a citation existence check at publication.
- Independent recomputation: the execution sandbox was unavailable, so a second language model recomputed 12 quantitative claims step by step in text from the paper's stated inputs (no code was executed); 0 differed from the stated value at the stated precision.
- Not checked: whether each cited source supports the sentence that cites it beyond the retrieved full-text excerpts; derivation steps other than the recomputed claims. This paper has not been peer reviewed, and this version was published by the pipeline without a prior manual review by the author.

References

[1] When Domains Collide: An Activity Theory Exploration of Cross-Disciplinary Collaboration. arXiv:2506.20063v4. https://arxiv.org/abs/2506.20063v4 [2] Toward Human-Centered AI-Assisted Terminology Work. arXiv:2512.18859v4. https://arxiv.org/abs/2512.18859v4 [3] Should Terminology Principles be re-examined?. arXiv:1609.05170v1. https://arxiv.org/abs/1609.05170v1 [4] Terminology-Aware Translation with Constrained Decoding and Large Language Model Prompting. arXiv:2310.05824v1. https://arxiv.org/abs/2310.05824v1 [5] Improved Multiscale Structural Mapping with Supervertex Vision Transformer for the Detection of Alzheimer's Disease Neurodegeneration. arXiv:2604.14837v1. https://arxiv.org/abs/2604.14837v1 [6] Evaluating LLM-driven User-Intent Formalization for Verification-Aware Languages. arXiv:2406.09757v2. https://arxiv.org/abs/2406.09757v2 [7] A White Box Analysis of ColBERT. arXiv:2012.09650v1. https://arxiv.org/abs/2012.09650v1 [8] What Generative Artificial Intelligence Means for Terminological Definitions. arXiv:2402.16139v3. https://arxiv.org/abs/2402.16139v3 [9] https://papers.qnfo.org/papers/terminology-silos-consilience. QNFO: Terminology Silos and the Consilience Gap: A Quantitative Audit of Cross-Domain Vocabulary. [10] https://papers.qnfo.org/papers/comprehensive-technical-framework-for-network-isomorphism. QNFO: Comprehensive Technical Framework for Network Isomorphism. [11] https://papers.qnfo.org/papers/cross-ratio-invariants-as-a-formal-measure-for-translation-fidelity-and-analogic. QNFO: Cross-Ratio Invariants as a Formal Measure for Translation Fidelity and Analogical Reasoning.