Knowledge Organization

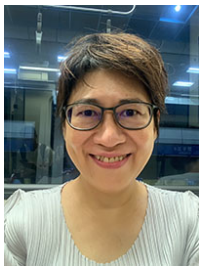
Knowl. Organ. **2026**; 53(5): 52642
https://doi.org/10.31083/KO52642

Article

SimRAG as Traceable Governance for Multilingual Thesaurus Localization: Evidence From AAT

Shu-Jiun Chen[1,2,*], Jeng-Shiun Wang[2]

[1]Institute of History and Philology, Academia Sinica, 11529 Taipei, Taiwan
[2]Academia Sinica Center for Digital Cultures, Academia Sinica, 11529 Taipei, Taiwan
*Correspondence: sophy@sinica.edu.tw (Shu-Jiun Chen)
Academic Editor: Natalia Tognoli
Submitted: 6 April 2026 Revised: 18 June 2026 Accepted: 7 July 2026 Published: 9 September 2026



Shu-Jiun Chen is an Associate Research Fellow at the Institute of History and Philology, Academia Sinica, and Executive Secretary of the Academia Sinica Center for Digital Cultures (ASCDC). She also holds academic appointments in library and information science at National Taiwan University and National Chengchi University. Her research interests include knowledge organization, cultural heritage informatics, digital humanities, linked data, digital libraries, and digital curation. Dr. Chen initiated the Chinese-language localization of the Art & Architecture Thesaurus (AAT) in collaboration with the Getty Research Institute and founded the Linked Open Data Lab at Academia Sinica. She is also known internationally as Sophy Chen.

Jeng-Shiun Wang holds a master's degree in Library and Information Science from National Taiwan University. He is Associate Editor of AAT-Taiwan, the Chinese-language version of the Getty Art & Architecture Thesaurus maintained by the Academia Sinica Center for Digital Cultures (ASCDC). His research interests include knowledge organization, controlled vocabularies, and information behavior. In his current role, he contributes to multilingual thesaurus localization and to research on the editorial use of controlled vocabularies in cultural heritage contexts. His master's thesis examined the information behavior of social-issue-focused YouTube creators.

Abstract

Multilingual thesauri support cross-lingual access and semantic interoperability, but localization remains difficult to scale because it requires stable concept boundaries, traceable decisions, and accountable maintenance. This exploratory design study treats multilingual thesaurus localization as both a translation problem and a knowledge organization governance problem. It proposes Simulated Retrieval-Augmented Generation (SimRAG) as a traceable design for examining artificial intelligence (AI)-assisted localization before a full retrieval-augmented generation environment is available. Using the Art & Architecture Thesaurus (AAT) as the empirical setting, the study examines 30 concepts across seven facets and focuses on two governance-critical tasks: term alignment and scope note translation. SimRAG constructs a controlled evidence setting using curated snippets from authoritative sources so that candidate terms, evidence links, risk notes, and editorial decisions can be reviewed together. In the internally adjudicated review setting, preferred-term status was stable across all 30 sampled concepts. Non-preferred-term alignment was more variable, with evidence-backed candidates showing higher observed alignment than model-inferred candidates, suggesting that candidate provenance is a useful cue for editorial prioritization. For scope notes, AI-drafted translations received higher adjudicated consensus ratings than existing translations on semantic accuracy, fluency, and cultural adaptation. These findings should be interpreted as design-oriented observations under controlled evidence conditions, without external validation of model quality. Qualitative analysis identified two recurring risks: semantic compression and pragmatic overextension. SimRAG's main contribution is a traceable editorial governance design that links evidence, candidate provenance, risk judgment, and knowledge organization system (KOS) status assignment. The paper contributes to knowledge organization by showing how AI-assisted multilingual thesaurus work can be organized as a reviewable, warrant-based decision process.

Keywords: SimRAG; multilingual thesaurus localization; traceable governance; editorial governance; candidate provenance; term alignment; scope note translation; warrant

1. Introduction

Multilingual thesauri and controlled vocabularies support cross-lingual access, semantic interoperability, and consistent indexing across communities and systems (Baca and Gill, 2015; Zeng and Chan, 2004). In galleries, libraries, archives, and museums (GLAM), they also function as shared semantic infrastructures for cataloging, discovery, and data integration. Yet multilingual thesaurus localization remains difficult to scale. The difficulty includes language, but it also involves stable concept boundaries, accountable editorial decisions, and long-term maintenance across languages and versions (Ménard et al., 2016). In knowledge organization (KO) terms, these requirements can be understood through warrant, that is, the justificatory basis for selecting and stabilizing labels and structures within a knowledge organization system (KOS) (Beghtol, 1995; Kwaśnik, 2010). Here, governance means a warrant-based, context-dependent, and traceable editorial decision process through which candidate labels, variants, and scope-note phrasings are justified, reviewed, assigned KOS status, and maintained over time. Recent advances in generative artificial intelligence (GenAI), especially large language models (LLMs), have renewed interest in semi-automating KO work, including multilingual drafting, terminology support, and vocabulary maintenance. In principle, LLMs can support multilingual drafting, paraphrasing, and terminology work at a scale that manual editorial processes cannot easily match. In practice, however, the use of LLMs in thesaurus localization raises two governance-critical concerns. First, AI-generated candidates and drafts must be traceable and reviewable. Editors need to know which evidence a model relied on and why a term or phrasing was suggested. Second, editorial review must be risk-aware. Not all concepts, term relations, or scope notes carry the same semantic or cultural risk. Treating all units as equally review-intensive weakens the practical value of AI assistance.

Retrieval-augmented generation (RAG) is often presented as a solution because it grounds generation in retrieved evidence. A full retrieval environment, however, involves more than technical infrastructure; it also requires curated corpora, retrieval design, evaluation procedures, and editorial rules. Many vocabulary programs do not have these conditions in place when they first begin to explore GenAI-assisted localization. This creates an editorial problem before it becomes a technical one: before a full RAG system is deployed, how can a vocabulary program establish a reproducible, evidence-grounded, and accountable human-AI design for experimentation and review?

This paper addresses that question by proposing Simulated Retrieval-Augmented Generation (SimRAG) as a traceable editorial governance design for multilingual thesaurus localization. SimRAG constructs a controlled evidence context by curating authoritative text snippets and aligning them to target concepts before generation. This design holds evidence conditions stable across cases and allows model outputs to be examined together with evidence links, risk flags, and editorial decisions. In this study, SimRAG names more than controlled-context prompting: it documents the connection among curated evidence, candidate provenance, risk notes, editorial review, and KOS status assignment. The study asks how a vocabulary program can establish an inspectable governance design before a full operational RAG environment is available, without treating the design as a causal test against a no-SimRAG baseline. We use the Art & Architecture Thesaurus (AAT) as the empirical setting and focus on two governance-critical editorial tasks: term alignment and scope note translation. Term alignment includes preferred and non-preferred terms and directly affects access points, synonym control, and retrieval behavior. Scope notes define concept boundaries, contextual limits, and culturally situated interpretations. These two tasks are central to thesaurus governance because both require explicit judgment about evidence, meaning, and acceptable use.

The study examines 30 AAT concepts across seven facets under a controlled evidence design. Two evidence tracks are used. The first examines term-alignment outcomes for preferred and non-preferred terms, with special attention to candidate provenance, that is, whether a candidate is supported by back-linkable evidence or inferred by the model. The second examines scope note translation through three Multidimensional Quality Metrics (MQM)-derived dimensions: semantic accuracy, fluency, and cultural adaptation. In addition, qualitative diagnosis is used to identify recurring governance-relevant risks in generated scope notes, especially semantic compression and pragmatic over-extension. The paper treats multilingual thesaurus localization as a governance problem in which evidence, generation, review, and decision records are explicitly linked. In this view, SimRAG is useful because it makes editorial decisions easier to inspect and reconstruct while supporting risk-tiered review. Under such a design, editorial attention can be organized more explicitly around high-risk units. A secondary and more limited concern is whether this governance design also shows bounded signs of review-stage workload redistribution during review and finalization.

This study is guided by three research questions.

RQ1: Within the SimRAG controlled evidence setting, what term-alignment patterns are observed for preferred and non-preferred terms, and how does candidate provenance relate to alignment outcomes and editorial review risk?

RQ2: In blinded author-editor consensus review, how do AI-drafted scope note translations compare with existing human translations in adjudicated ratings of semantic accuracy, fluency, and cultural adaptation?

RQ3: As a supporting question, what signs of bounded review-stage workload redistribution are observed when AI drafting is followed by editor verification?

This paper makes three contributions. First, it proposes SimRAG as a traceable, warrant-based editorial governance design for multilingual thesaurus localization before full RAG deployment. Second, it examines term alignment and scope note translation within a shared governance framework, with MQM-based evaluation applied specifically to scope note translation. Third, it derives a practical risk-tiered review approach in which provenance and qualitative risk diagnosis support focused editorial attention and candidate-to-status decision records. The workload observations are limited to review-stage redistribution; full process cost is outside the study scope.

The remainder of the paper is organized as follows. Section 2 reviews prior work on multilingual thesaurus governance, generative AI in knowledge organization, traceable generation, and multidimensional quality assessment. Section 3 presents the research design, including sampling, evidence package construction, prompting strategies, and evaluation procedures. Section 4 reports the quantitative and qualitative results. Section 5 discusses governance implications, risk-tier review rules, and the broader relevance of the editorial design. Section 6 concludes with contributions, limitations, and directions for future work.

2. Literature Review

2.1 Multilingual KOS and Thesaurus Localization: Interoperability, Term Control, and Scope Notes

Multilingual KOS aim to support cross-lingual retrieval and semantic interoperability. A persistent challenge is maintaining conceptual consistency across languages while accommodating differences in linguistic form, disciplinary practice, and cultural framing. Interoperability is therefore not limited to bilingual term mapping; it also concerns concept modeling, relationship representation, structural and semantic compatibility, and governance arrangements among vocabularies that may differ in structure, domain, language, or granularity (Zeng and Mayr, 2019). For thesauri, preferred and non-preferred terms provide controlled access points and support synonym management, while scope notes define concept boundaries and conditions of use. Scope notes are a key mechanism for preventing semantic confusion and long-term structural drift, particularly when maintenance is distributed across teams and time (Getty Vocabulary Program, 2024).

Cross-language thesaurus work commonly faces two coupled issues. First, semantic alignment is difficult when concepts are culture loaded or domain specific, and when the target language community has distinct naming practices and interpretive conventions. Hjørland (2007) emphasizes that semantic organization is not a neutral linguistic conversion, but is shaped by classificatory purposes, community norms, and epistemic perspectives. This implies that alignment decisions must incorporate community acceptability, not merely lexical similarity. This issue is closely related to warrant, because term choice and structural fit must be justified not only by lexical equivalence but also by domain evidence and by the epistemic commitments of the community for which the vocabulary is designed (Bullard, 2017; Martínez-Ávila and Budd, 2017). Second, long-term maintenance requires consistent editorial rules and traceable decision records. Without governance mechanisms that document rationales and edits, concept boundaries can drift, weakening structural consistency, accountability, and reproducibility.

Domain analysis further clarifies why semantic alignment cannot be treated as lexical substitution. From a domain-analytic perspective, knowledge organization should be examined in relation to the purposes, documents, terminologies, and discourse practices of specific domains (Hjørland, 2002). Multilingual thesaurus work must therefore ask whether a candidate expression fits not only the source concept's scope-note and hierarchy-defined boundary, but also the disciplinary usage, target-language convention, and epistemic commitments of the community for which the vocabulary is maintained. Cultural warrant adds an ethical and representational constraint: Beghtol (2002) argues that knowledge representation and organization systems should be culturally acceptable and hospitable. In the present context, localized labels and scope notes should therefore be usable in the target-language knowledge environment without authorizing locally familiar expressions that distort the source concept's institutional, historical, or classificatory constraints. For this reason, acceptability in this study refers to editorial acceptability for KOS status assignment: whether a candidate can responsibly be authorized as a preferred term, retained as a non-preferred access point, treated as a contextual variant, or rejected.

Standards and large-scale infrastructures reinforce the need to integrate conceptual modeling, data exchange, and governance. ISO 25964-1 provides guidance for thesaurus construction and maintenance and highlights standardized data structures for interoperability (International Organization for Standardization, 2011). In the semantic web environment, Simple Knowledge Organization System (SKOS) provides a shared representation model to publish, exchange, and link thesaurus data across systems (Miles and Bechhofer, 2009). Practical implementations further show that governance arrangements are essential for sustainable quality control. AGROVOC, a multilingual agricultural thesaurus, aligns multilingual KOS in SKOS with expert validation as a core quality mechanism (Caracciolo et al., 2012). EuroVoc emphasizes cross-institutional coordination supported by collaborative tools (Walhain et al., 2025). VocBench highlights the importance of role differentiation and workflow formalization for maintaining authoritative vocabularies (Stellato et al., 2015). Within Chinese language contexts, work on art vocabularies has shown that semantic, structural, and cultural differences jointly shape alignment outcomes (Chen et al., 2016). Together, these standards, infrastructures, and studies motivate treating

term alignment and scope note translation as governance-critical work units that can be defined, evaluated, and maintained under explicit rules.

2.2 Generative AI in Knowledge Organization: From Output Production to Governed Decision Support

Recent discussion of generative AI in knowledge organization has shifted from content production toward the conditions under which AI can support editorial decision-making. For vocabulary and thesaurus work, speed and output volume are insufficient if generated candidates and drafts cannot be traced, reviewed, and incorporated into accountable maintenance decisions. In multilingual and semantic resource settings, recent work on LLM-aided SKOS thesaurus translation suggests that LLMs can support structured multilingual processing, while still requiring traceable evidence and human review to reduce risks of boundary misplacement and structural inconsistency (Kraus et al., 2025). Even as multilingual model capability improves (Yan et al., 2024), core KO challenges remain centered on conceptual equivalence, boundary maintenance, and governance of cultural and pragmatic constraints. This suggests a role for LLMs as decision support that proposes candidates, articulates reasons, and flags risks, while preserving editorial authority and responsibility in human governance.

Methodologically, responsible adoption of generative AI in KO tasks must address three requirements. First, system outputs should be aligned to governance-relevant units, such as preferred and non-preferred term alignment and scope note drafting. Second, evaluation must be designed for reproducibility to support cross-version comparison and risk diagnosis. Third, human and AI division of labor should be organized through risk tiering so that review effort concentrates on high-risk units rather than scaling linearly with output volume. These requirements provide the rationale for framing thesaurus localization as a governed editorial process rather than a translation problem.

2.3 Retrieval-Augmented and Traceable Generation: Why SimRAG Matters as a Research Design

RAG aims to reduce hallucinations and outdated knowledge by grounding generation in externally retrieved evidence. The RAG architecture links a retriever to an external index so that generation can refer to specific contexts, aligning with governance expectations for traceability and accountability in KOS maintenance (Lewis et al., 2020). However, later surveys show that RAG performance depends strongly on retrieval quality, context construction, and alignment strategy. When retrieval-stage variables are not controlled, output differences become difficult to attribute, which limits reproducible evaluation and comparison (Gao et al., 2024). Self-RAG further argues for selective retrieval and post-generation self-reflection to reduce instability caused by irrelevant contexts, reinforcing the need for explicit criteria and controls in the retrieval process (Asai et al., 2024).

A complementary line of work addresses faithfulness and traceability through attribution frameworks and evidence-grounded generation. Rashkin et al. (2023) conceptualize and evaluate attribution in natural language generation. Tianyu Gao et al. (2023b) benchmark citation-grounded answer generation, while Luyu Gao et al. (2023a) show how generated text can be revised after generation to improve attribution and evidential support. Slobodkin et al. (2024) further develop locally attributable grounded generation. Because retrieval noise and retriever limitations can distort evaluation, controlled designs and verification-oriented strategies remain important, especially for terminological tasks in which boundary control is central.

In this study, the related terms are distinguished as follows. Provenance identifies the source or origin of a candidate term, definition, or evidence snippet. Evidence grounding indicates whether an AI-generated candidate or draft can be linked back to such evidence. Traceability refers to a broader decision-level property: the ability to reconstruct how evidence, AI output, risk judgment, editorial review, and final KOS status were connected in a decision record. This distinction is important for cultural heritage vocabularies, where semantic interoperability depends not only on shared labels but also on the preservation of interpretive context. Work on CIDOC Conceptual Reference Model (CIDOC CRM) similarly shows that integration across cultural heritage data depends on explicit conceptual relations and the conditions under which information is produced and reused (Doerr, 2003).

These results motivate SimRAG as a practical interim design before full RAG deployment. When retrieval infrastructure is not yet available, manually curated controlled contexts can serve as a stable evidence setting rather than as a claim to full retrieval. Holding evidence conditions stable supports concept-by-concept review of term alignment and scope note drafting, while also making the limits of the evidence setting visible for later operational RAG implementations (Gao et al., 2024; Lewis et al., 2020).

SimRAG can be distinguished from related forms of AI assistance. Controlled-context prompting supplies background context to the model, but controlled context alone does not specify how candidate status, risk, and editorial decisions are recorded (Gao et al., 2024; Lewis et al., 2020). A general human-in-the-loop process keeps people in the review process, but human participation does not by itself make decision criteria or KOS status changes explicit. Evidence-based LLM-assisted translation uses sources to support a translation, but its main concern is usually textual quality and attribution rather than controlled-vocabulary authorization (Kraus et al., 2025; Rashkin et al., 2023). SimRAG combines these elements in a narrower KO sense: each model suggestion is treated as a candidate to be linked to provenance, checked against scope-note and hierarchy constraints, assigned a review risk, and then accepted, downgraded, annotated, or rejected as a KOS decision.

2.4 Multidimensional Quality Assessment and Reliability: MQM Configuration and Bias Control

Translation quality assessment must support diagnosis and governance actions rather than rely on single overall impressions. The MQM framework provides a hierarchical typology of issue categories that enables quality to be decomposed into observable and annotatable dimensions across contexts (Lommel et al., 2013). MQM is configurable, allowing evaluators to select dimensions aligned with task requirements and use settings (Lommel et al., 2013). Related work stresses that evaluation dimensions should be tightly aligned with task specifications (Mariana et al., 2015). Cross-lingual applications also indicate that category definitions and workflows may require language-aware adaptation. For East Asian languages, direct use of typologies optimized for European languages can lead to ambiguous category boundaries, motivating adaptations and the use of consistency checking as a validation component (Silva et al., 2024).

Reliability is central when multidimensional assessment is used for scholarly claims and governance decisions. Fine-grained annotation can be diagnostically useful but often suffers from low agreement due to ambiguity in span definition, category boundaries, and attribution judgments. Clear guidelines, training, calibration, and consistency checks are therefore necessary (Lommel et al., 2014). Agreement is commonly measured using Cohen's kappa or Fleiss' kappa, complemented by consensus discussions (Cohen, 1960; Fleiss, 1971). Reliability is also affected by evaluation unit choices and the provision of context. Sentence-level and document-level judgments involve trade-offs among agreement, cost, and cognitive burden, and context presentation can change outcomes, so transparent reporting of evaluation units and calibration procedures is important for reproducibility (Knowles and Lo, 2024).

LLM-as-a-judge approaches can reduce cost but introduce additional risks. Studies report sensitivity to prompts and systematic deviations from human judgment (Chen et al., 2024; Koo et al., 2024). Pairwise comparison designs require attention to position bias and repeated stability checks (Shi et al., 2025). Self-preference bias may also occur when model judges favor outputs closer to their own style distributions (Wataoka et al., 2024). These findings support positioning LLM-based assessment as structured pre-screening and diagnostic support rather than as a final arbiter, and motivate mitigation procedures such as blinding, calibration, and consistency checks to reduce drift and bias (Chen et al., 2024; Koo et al., 2024; Lommel et al., 2014).

2.5 Cultural Adaptation and Culture-Loaded Terms: Governance Risks in Definitional Thesaurus Texts

In multilingual thesaurus localization, cultural adaptation concerns whether terms and scope notes can be correctly understood and applied in the target-language knowledge environment. For thesauri, scope notes function as definitional texts that delimit meaning and constrain usage. Misplaced cultural contexts or inappropriate analogies can shift concept boundaries, affecting hierarchical placement and alignment stability and reducing retrieval usability (Getty Vocabulary Program, 2024; Zeng and Mayr, 2019). Culture-loaded terms raise higher risks because their meanings depend on culturally specific knowledge. Such issues have been documented in Chinese-language art vocabulary alignment where semantic scope and cultural framing jointly influence outcomes (Chen et al., 2016).

From a KO perspective, cultural adaptation reaches beyond natural-sounding translation. Naming and classification are interpretive practices that authorize some access paths while limiting others (Olson, 2002). In a thesaurus, a preferred term does not merely label a concept; it stabilizes one form of representation as central, while non-preferred and rejected terms shape the surrounding space of retrieval and interpretation. Trustworthy classification also depends on transparency about the grounds of decisions and about the biases that may enter classification practice (Mai, 2010). For this reason, cultural adaptation is treated here as a governance dimension: it asks whether a localized term or scope note preserves the classificatory role of the AAT concept, remains usable in the target-language community, and avoids representational risks such as semantic over-breadth, cultural misplacement, or misleading register.

Translation studies describe culture-specific items as expressions without stable equivalents in the target culture. Literal translation can misplace meaning, while excessive localization can erase institutional constraints and historical context (Aixelá, 1999). Strategy selection is therefore best understood as functional and context-dependent decision-making evaluated against communicative purposes and audiences, rather than as rule-like substitution (Molina and Hurtado Albir, 2002). For thesaurus scope notes, the governance implication is that definitional translation should preserve institutional and historical constraints. When entries involve institutions, rituals, craft processes, or art-historical classificatory traditions, intuitive near-equivalents can introduce semantic drift with downstream effects on synonym control and cross-language alignment governance. Cultural adaptation can capture cross-cultural risks that semantic accuracy alone may not reveal, supporting its treatment as an independent governance-relevant dimension (Harpring, 2010).

Recent machine translation and LLM research similarly highlights cultural context as a key variable. Contextual translation challenges include cultural details, idioms, and domain terminology (Naveen and Trojovský, 2024). Cultural competence has been measured and benchmarked in evaluation settings (Yao et al., 2024), while studies also report that fluent outputs may still fail to balance comprehensibility and local acceptability for cultural details (Budimir, 2025; Singh et al., 2024). Despite these ad-

vances, definitional texts in controlled vocabularies have received less attention than general texts, leaving a methodological gap in how cultural risks should be evaluated and governed for thesaurus maintenance. This gap motivates treating cultural adaptation as a diagnosable risk dimension for definitional thesaurus translation and feeding documented deviations back into traceable maintenance workflows (Getty Vocabulary Program, 2024; Zeng and Mayr, 2019).

3. Research Design and Implementation

3.1 Empirical Setting and Sampling Strategy

This study uses the AAT as an empirical setting to examine governed human and AI collaboration in multilingual thesaurus localization. The design treats two task units as governance critical for long-term vocabulary maintenance: term alignment (preferred and non-preferred terms) and scope note translation. To observe alignment and risk profiles under varying conceptual granularity and cultural context, we adopted stratified sampling rather than random sampling. Thirty representative concepts were selected across the seven AAT facets (the full list is provided in Appendix A). The stratification criteria were hierarchical level and cultural reference, with the intent to ensure sufficient coverage of these two variables for comparative analysis. We limited the sample to 30 concepts to enable concept-by-concept qualitative adjudication; the author-editors typically discussed each concept for four hours or more across the alignment and scope note review stages. The 30-concept sample was designed to support concept-level adjudication and design-oriented observation rather than population-level statistical inference. The stratification was used to ensure variation in hierarchical granularity and cultural reference, both of which may affect boundary maintenance and editorial risk.

Hierarchical coverage was defined using the AAT hierarchy. Upper-level concepts were defined as those within the first two hierarchical levels (facet and sub-facet levels). Lower-level concepts were defined as those from the third level onward. This distinction allows the study to examine whether localization outcomes differ with conceptual specificity and granularity. Hierarchical level was therefore treated as a sampling variable rather than as a basis for broad statistical inference. The distinction was intended to help observe whether broader concepts, whose labels often function as general access points, and more specific concepts, whose labels may depend more strongly on domain conventions and scope-note boundaries, present different kinds of editorial risk.

Cultural reference was assigned by examining the AAT preferred term, scope note, hierarchical placement, and domain use of each concept. East Asia-related concepts are those whose interpretation depends primarily on East Asian cultural, historical, religious, artistic, or material traditions. Western-related concepts are those whose interpretation depends primarily on Western, European, Mediterranean, or related art-historical traditions. Culture-general concepts are those whose basic interpretation does not depend on one specific cultural tradition, although they may still require local editorial choices in Traditional Chinese.

Editorial acceptability is introduced here as a review variable and applied in the term-alignment evaluation. It refers to whether a candidate expression can be responsibly assigned a KOS status under AAT scope-note and hierarchy constraints in the target-language setting. This criterion differs from general linguistic acceptability: a fluent candidate may still be unacceptable if it broadens the concept boundary, shifts the cultural frame, conflicts with target-locale usage, or weakens retrieval guidance.

Task alignment was ensured by using the same set of thirty concepts for both sub-tasks. This provides a shared conceptual reference set for cross-task comparisons, avoiding confounding introduced by using different concept sets for term alignment and scope note translation. Table 1 summarizes the resulting distribution of the 30 sampled concepts by hierarchical level and cultural reference. These variables are used to structure case selection and review, not to support population-level generalization. The domain of analysis is more limited than AAT localization in general: English-to-Traditional-Chinese AAT localization under controlled evidence conditions, examined through two reviewable editorial task units.

3.2 Data Sources and Evidence Package Construction

To support traceability and accountable review, we constructed an evidence package with two roles: (1) providing controlled context for SimRAG in the term alignment task, and (2) serving as the primary back-linking resource for editor verification and subsequent risk analysis.

The evidence package is treated here as a controlled evidence setting. It was curated before generation to make the evidential basis of each model suggestion inspectable and comparable across concepts. This design improves reviewability, while also creating a more favorable and less noisy setting than many live retrieval environments. The results therefore reflect this curated evidence condition and do not represent outcomes from an open or fully deployed retrieval setting.

The package integrates two evidence layers. The first is AAT English source data for each concept, including the preferred term, scope note, and hierarchical context (facet membership and broader-term context), which together define intended meaning and usage boundaries. The second layer comprises authoritative reference sources with stable citations to support cross-language comparison of definitions and domain notes, including the National Academy for Educational Research (NAER) Web of Words (LeCi-Wang; https://terms.naer.edu.tw/), Cambridge Dictionary, Collins Online Dictionary, and Wikidata.

Table 1. Distribution of 30 concepts by hierarchical level and cultural reference.

Cultural reference	Upper level (first two levels)	Lower level (third level onward)	Total
East Asia related	2	5	7
Western related	0	10	10
Culture-general	10	3	13
Total	12	18	30

All materials were consolidated into a structured table with fields for the English source term, AAT scope note, candidate Traditional Chinese alignments, reference definitions or domain notes, and source links. During experiments, the evidence package was embedded into prompts as structured text snippets, simulating retrieval outputs prior to deploying a full RAG infrastructure. This design allows reviewers to trace model judgments to the same curated evidence used at generation time and to compare AI outputs with existing human translations and editorial decisions.

Candidate provenance was recorded as a review cue. Evidence-backed candidates were terms drawn from or directly supported by the curated evidence package or the authoritative reference sources. Model-inferred candidates were terms proposed by the model without direct support in the evidence package. This distinction did not automatically determine acceptance or rejection; it identified candidates requiring closer editorial review.

Because the target language is Traditional Chinese, candidate terms and scope note drafts are recorded in Chinese characters. For international readability and reuse, concept identifiers, English labels, and evidence links function as the primary indexing keys, enabling readers who do not read Chinese to follow the evidence trail and interpret decisions at the concept level.

3.3 Models, Task Decomposition, and Prompting Strategies

The study comprises two sub-tasks, term alignment and scope note translation, and applies differentiated model and prompting strategies consistent with task decomposition principles. Term alignment emphasizes conceptual equivalence judgment and candidate comparison across multiple options. Scope note translation emphasizes definitional writing quality in Chinese while preserving conceptual boundaries and readability. Full prompts and worked examples are provided in Appendix B.

The use of different models was a task-bound design choice, not a model-comparison design. At the time of experimentation, OpenAI o3 and GPT-4o, both accessed through ChatGPT, offered complementary affordances for the two editorial tasks. Model selection was therefore guided by task fit and preliminary internal task checks rather than by an experimental comparison of model performance. Term alignment required reasoning-intensive comparison among candidate expressions, provenance checking, and risk explanation. Scope note translation required readable Traditional Chinese definitional prose while preserving the source concept's scope-note boundary. Because the two tasks used different models, the study cannot separate task effects from model-specific effects, and the findings apply only to the selected model-task conditions.

3.3.1 Term Alignment With SimRAG Controlled Context

Term alignment experiments used OpenAI o3 (snapshot 2025-04-16; OpenAI, San Francisco, CA, USA). The testing period was April to June 2025. Prompts implemented SimRAG by embedding controlled context constructed from the evidence package. For each concept, the prompt provided the AAT scope note, a set of candidate Chinese terms (hereafter candidates), and authoritative reference definitions or notes in a structured format. The model was instructed to compare candidates semantically, judge conceptual equivalence, and produce reviewable rationales and risk flags.

OpenAI o3 was used for this task because term alignment required comparative judgment among candidate expressions, attention to evidence links, and explicit risk explanation, all of which were central to candidate-status review. This choice does not make OpenAI o3 uniquely suited to thesaurus term alignment; other models may produce different candidates, rationales, or risk notes.

The rationale for this design is methodological and governance oriented. In many vocabulary programs, a full retrieval system is not yet available when experimentation begins. SimRAG allows the study to observe model behavior for equivalence judgment and synonym control under controlled evidence conditions, while producing signals that can support professional editorial decision-making.

The required output for each concept included five elements. First, the model selected one preferred term. Second, it recommended a set of non-preferred terms. Third, it listed terms not recommended for adoption and explained semantic or pragmatic risks. Fourth, any additional candidate terms proposed by the model itself were explicitly labeled as model-proposed terms, distinguishing them from evidence-derived candidates. Fifth, the model stated its judgment criteria and provided back-links to the relevant evidence snippets on which it relied. This output design distinguishes between non-preferred terms and disfavored terms to improve screening clarity and to preserve traceability at the decision level.

In this study, the target locale for preferred labels is Traditional Chinese as used in Taiwan, reflecting the in-

tended deployment context and editorial conventions of the localization program. Regional lexical variants are treated as variants that may be recorded as non-preferred labels or notes when they support retrieval, but they are not used as the default preferred forms under the defined governance scope.

3.3.2 Scope Note Translation Under a Zero-Shot Communicative Strategy

Scope note translation experiments used GPT-4o (snapshot 2025-01-29; OpenAI, San Francisco, CA, USA). The testing period was April to June 2025. To establish a baseline for evaluation, we adopted a zero-shot strategy. The model received only task background and translation principles, without examples or additional in-context training data. This design observes drafting behavior under a zero-shot communicative instruction and does not examine imitation of provided exemplars.

GPT-4o was used because the task required fluent Traditional Chinese definitional prose while preserving the conceptual boundaries of the English scope notes. Preliminary internal task checks suggested that GPT-4o produced more readable scope-note drafts than OpenAI o3 in this setting. This observation was used only as practical task-selection guidance and was not designed, reported, or interpreted as a model-comparison experiment. Because a different model was used for this task, the observed ratings apply only to this model-task setting and do not generalize to other models.

Two principles guided scope note translation. The first is fidelity. The translation must accurately convey the intended meaning and definitional boundary conditions of the English scope note. The second is a communicative translation orientation. Without altering the conceptual core, the translation should prefer expressions that are understandable to target-language readers and consistent with appropriate register norms for definitional text. Separately, prompt-level pilot testing indicated that explicitly specifying a communicative orientation helped reduce unnatural literal phrasing and improve readability while retaining the objective tone expected of scope notes.

3.4 Procedure

The procedure comprises three phases: data preparation, generation experiments, and evaluation (Fig. 1). Steps 2 to 4 apply only to the term alignment task because they build the SimRAG controlled context. Scope note translation proceeds directly to the zero-shot experiment.

Phase 1 prepares the sample, candidate lists, and the evidence package for SimRAG (Steps 2–4 apply only to term alignment). Phase 2 generates term-alignment drafts and zero-shot scope note drafts. Phase 3 conducts editorial verification for term alignment and MQM-based review for scope note translation.

3.5 Exploratory Evaluation and Editorial Adjudication

The evaluation should be understood as an exploratory author-editor adjudication study. The two evaluators were also the authors and long-term AAT Traditional Chinese editors. This expertise allowed judgments to be grounded in established AAT localization practice, but it also limits evaluator independence. Consensus adjudication was used to create stable decision records for governance analysis, not to estimate how independent external raters would score the same outputs. Because the sample is small and the final results are consensus outcomes, the quantitative results are treated as design-oriented summaries under the present review setting.

3.5.1 Term Alignment

Term alignment was evaluated as editorial acceptability for KOS status assignment. Two author-editors judged whether each candidate could be responsibly accepted under AAT scope-note and hierarchy constraints in the Traditional Chinese target setting. Acceptable candidates preserved the concept boundary, fit the relevant hierarchy, and had evidence support or a defensible editorial rationale. Unacceptable candidates were those that introduced semantic over-breadth, cross-domain drift, cultural misplacement, misleading register, or weak target-locale usability. Judgments were anchored to AAT editorial practice rather than to general subject-matter preference.

For reporting, we calculated a concept-level macro-average alignment measure. Because the number of candidates differed across concepts, each concept was first summarized by the proportion of acceptable candidates and then all thirty concept-level proportions were averaged equally. This measure prevents concepts with many candidate terms from dominating the result, but it should be interpreted as an exploratory summary of adjudicated editorial judgments rather than as a population-level estimate. For this statistic, “non-preferred” refers only to candidates proposed for non-preferred-term status. Disfavored candidates were adjudicated separately on whether the proposed exclusion was upheld and were not included in the non-preferred-term macro-average.

We also compared existing human labels and AI outputs qualitatively to identify governance-relevant bias patterns, especially semantic drift and pragmatic risk. The two editors first judged independently and recorded reasons. Initial pre-consensus agreement was 96 of 106 candidate-level judgments (90.6%) overall, including 27 of 30 preferred-term judgments (90.0%) and 69 of 76 non-preferred-term judgments (90.8%). Most disagreements clustered in boundary-sensitive or culturally layered cases and were resolved through adjudication with reference to scope notes, hierarchy, and additional sources. All cases were then discussed in editorial meetings until a single adjudicated decision record was reached for each item. The reported pre-consensus agreement figures describe the review

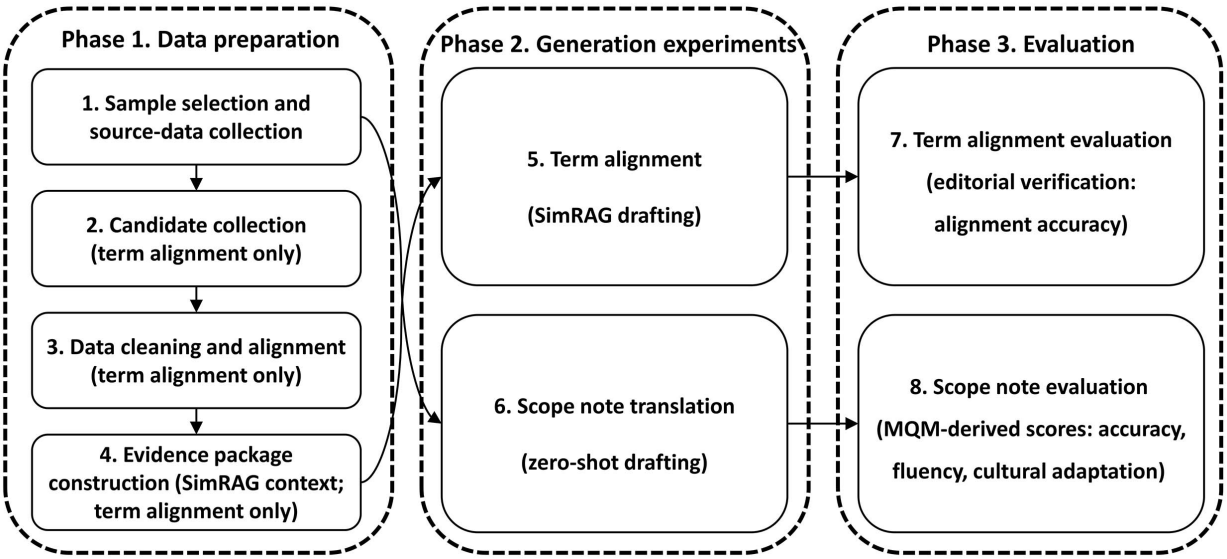


Fig. 1. Procedure: data preparation, generation experiments, and evaluation (Steps 2–4 apply only to term alignment). SimRAG, Simulated Retrieval-Augmented Generation; MQM, Multidimensional Quality Metrics.

process; the final analytical records are consensus records and are therefore not statistically independent rater scores.

3.5.2 Scope Note Translation

Scope note translation quality was assessed with an MQM-based scheme configured for definitional thesaurus text. The assessment was treated as exploratory consensus review; it did not provide external validation. Three dimensions were used: semantic accuracy, fluency, and cultural adaptation. Appendix C provides the MQM-derived rubric used in model-assisted pre-assessment, including dimension labels, error categories, and scoring anchors. A worked example of the model-assisted pre-assessment procedure is provided in **Supplementary Material**. Cultural adaptation covered locale conventions, culture-specific references, and register. Each dimension was rated on a five-point Likert scale.

Evaluation proceeded in three stages. First, translations were blinded and randomized as Version A and Version B. Second, a language model applied the rubric for error marking and provisional scoring as diagnostic support only. To reduce position bias, we used symmetric prompts and reversed the A/B order for a subset of items. Third, two author-editors independently scored each version and then discussed all items in editorial meetings until a single adjudicated consensus score was reached. At the binary screening stage used before adjudication, initial agreement was 86 of 90 dimension-level judgments (95.6%) overall: 27 of 30 for semantic accuracy, 30 of 30 for fluency, and 29 of 30 for cultural adaptation. These figures describe the pre-consensus screening judgments and should be distinguished from the final five-point consensus ratings used in the statistical analyses below. Adjudication notes and revision rationales were retained for later analysis of typical error patterns and cultural risks.

Because final ratings were reached through author-editor consensus rather than independent external evaluation, they may reduce observed variability. This is especially relevant to dimensions where translations were judged uniformly high, such as fluency, and may contribute to ceiling effects. Accordingly, the paired tests in Section 4 are reported as exploratory summaries of observed consensus differences rather than as external validation of model quality.

3.6 Limitations and Threats to Validity

Several limitations define the scope of inference. First, the study used a controlled evidence setting. It did not test a live retrieval environment. Curated evidence made the evidential basis of model suggestions inspectable and comparable across concepts, but it may also have created more favorable and less noisy conditions than many real vocabulary projects. The findings therefore are observations under curated evidence conditions and do not support a causal claim over a no-SimRAG baseline, alternative prompting strategies, or later operational retrieval settings.

Second, the evaluation was an author-editor adjudication study rather than an independent external evaluation. The author-editors' long-term experience with AAT Traditional Chinese localization grounded the judgments in established editorial practice, but the evaluators were also the authors. This limits evaluator independence. The absence of independent external evaluators means that the results should be read as internally adjudicated editorial findings rather than as independently validated quality judgments.

Third, consensus adjudication and non-independent ratings affect interpretation. Consensus was appropriate for producing stable decision records for governance analysis, but it also reduces observed variability. The final consensus scores are not statistically independent rater scores. Findings such as stable preferred-term alignment or uniformly high fluency scores should therefore be interpreted cautiously, especially where ceiling effects may occur. The initial agreement figures reported in Section 3.5 describe the pre-consensus review process, whereas the analyses in Section 4 use adjudicated consensus outcomes.

Fourth, the small sample limits statistical inference. The thirty concepts were selected to support concept-level qualitative adjudication across facets, hierarchy levels, and cultural reference types. They do not provide a basis for population-level estimates of alignment or translation quality across AAT or other thesauri. The paired tests and effect sizes reported below are therefore exploratory summaries of observed consensus differences within this sample, not robust empirical validation of model quality.

Fifth, model choice is a boundary condition. Term alignment and scope note translation used different models because they were treated as different editorial tasks. This design cannot separate task effects from model-specific effects. The findings should therefore be interpreted as observations under the selected model-task pairings, and later studies should examine whether similar patterns hold across other models, versions, and prompting conditions.

Sixth, the workload comparison is limited in scope. The reported time comparison concerns only review and finalization after evidence preparation and AI drafting. It excludes evidence package construction, prompt preparation, and the longer-term work required to maintain authoritative sources and editorial rules. This is a bounded observation about review-stage workload redistribution; the study does not estimate full cost reduction.

Finally, external validity is limited by the empirical setting. The study concerns English-to-Traditional-Chinese localization of AAT concepts. Concept types, scope-note conventions, target-locale usage, and reference-source characteristics may have shaped the observed patterns. Transfer to other thesauri, domains, or language pairs requires stable definitional texts, authoritative evidence, explicit editorial rules, and a risk-tier rubric for review prioritization. Without these conditions, comparable levels of accountability and reproducibility may not be achieved. Results in Section 4 should be interpreted within these boundaries, as design-oriented observations about traceable governance rather than as general validation of AI-assisted localization quality.

4. Results

4.1 Term Alignment Outcomes

4.1.1 Observed Alignment Patterns and the Governance Value of Traceable Evidence

The following results describe adjudicated term-alignment outcomes under the selected model and controlled evidence setting. Across the 30 sampled concepts, preferred-term alignment was stable: the author-editors accepted the preferred-term status for all sampled concepts after consensus review. Within this controlled sample, the preferred-term task was therefore less ambiguous than non-preferred-term governance.

Non-preferred-term alignment was more variable. The adjudicated macro-average was 71.9%, reflecting additional judgments about semantic extension, target-locale usage, and pragmatic constraints. As shown in Fig. 2, evidence-backed candidates showed higher observed alignment than model-inferred candidates (83.7% vs. 61.4%). This provenance-related difference functions as a review cue: candidates without direct evidence support may require closer editorial review.

4.1.2 Review-Stage Workload Redistribution Under SimRAG

In conventional thesaurus localization, editors often cross-check reference definitions or domain notes, scope notes, and hierarchical context to judge candidate fit and cultural appropriateness. At scale, this review work can become difficult to sustain. In this study, SimRAG consolidated scope-note context and authoritative references into a controlled prompt and required reviewable rationales and risk notes. This did not remove verification; it reorganized the review stage by making the evidence basis and possible risks easier to inspect.

Process logs suggest a bounded comparison of about 150 minutes per concept in the traditional process and about 10 minutes in the review-and-finalization stage under AI drafting plus editor verification. The 150-minute estimate covers candidate collection, definition checking, scope-note and hierarchy comparison, and final editorialization. The 10-minute estimate covers review, correction when needed, and finalization after AI drafting. It excludes evidence package construction, prompt preparation, source selection, and longer-term maintenance of editorial evidence. The comparison therefore characterizes review-stage redistribution under controlled conditions, rather than full-process cost.

4.1.3 Decision Support Signals and an Illustrative Boundary Case

In about one-third of the 30 concepts, the model suggested a preferred label different from the existing human-preferred label. After editorial review, several of these alternatives were judged to fit the scope note and hierarchical placement more closely. This indicates that model outputs

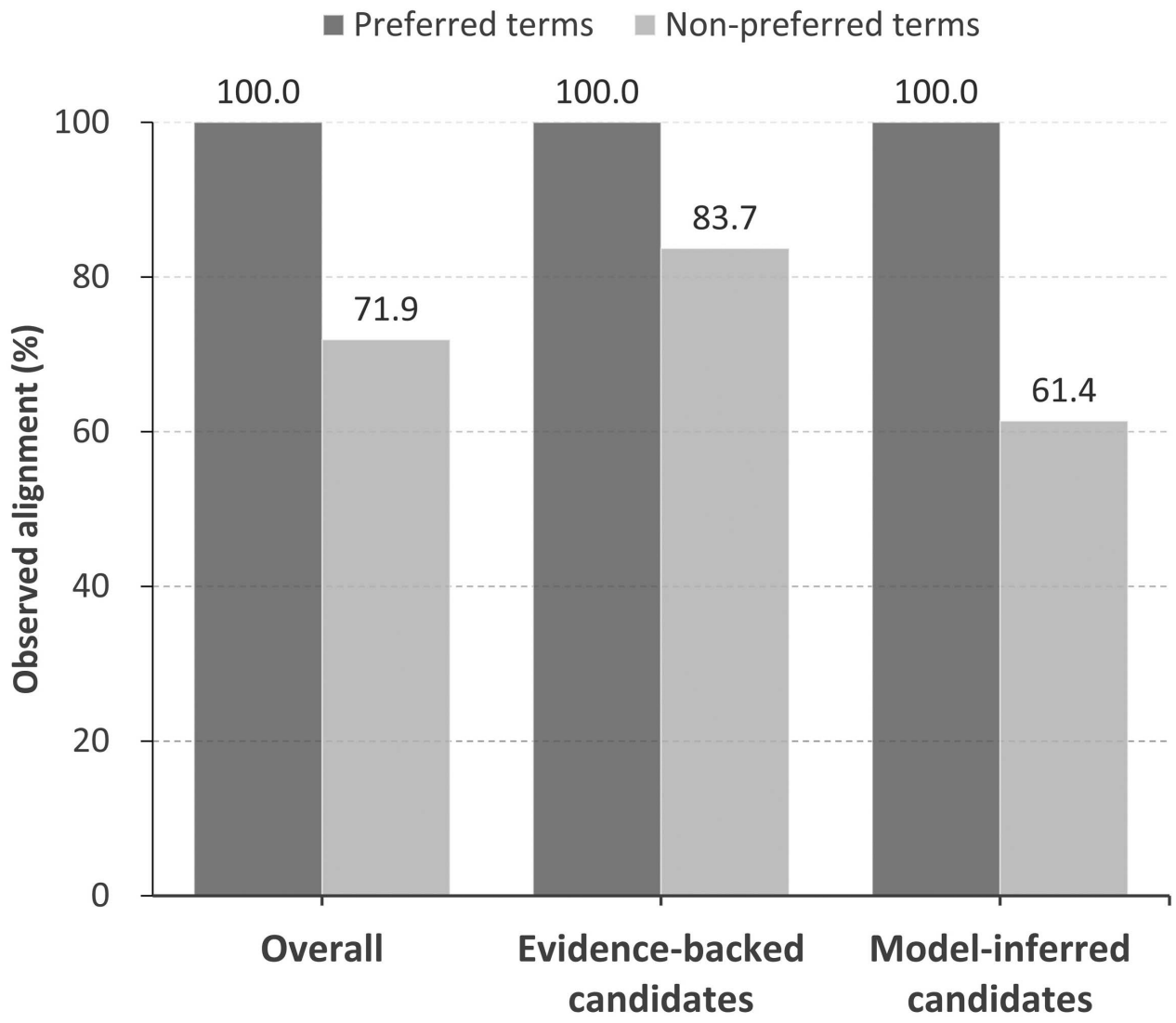


Fig. 2. Observed alignment for preferred and non-preferred terms, overall and stratified by candidate provenance. Note. N = 30 concepts. Alignment is computed as a concept-level macro-average, giving equal weight to each concept. The provenance-related difference is used as a review cue for editorial prioritization within this controlled sample.

can serve as review prompts rather than substitutes for editorial authority. Their practical value lies in drawing attention to boundary-sensitive cases, such as sense misplacement and cross-domain drift, so that editors can examine decisions that require conceptual trade-offs.

A representative case is motifs (AAT ID 300009700). In art-historical discourse, motifs may refer either to a decorative unit or to a recurring symbolic theme. In the existing human label set, the preferred label reflected the thematic sense, whereas the AI suggested a label for the decorative-unit sense. In AAT, the entry "motifs" appears in a Design Elements context, and its scope note defines a concrete decorative unit rather than iconographic or narrative symbolism.

Here, SimRAG linked candidate terms, scope-note constraints, and authoritative references, allowing editors to verify the model's reasons against thesaurus semantics. The case illustrates candidate-to-status governance rather than simple correction of a translation. In the decision record, each candidate received a proposed status—preferred, non-preferred, or disfavored—together with provenance and a review rationale; the proposal could then be upheld or rejected during editorial review. This structure supports review of a wider synonym and near-synonym space without treating every generated string as equally adoptable. Table 2 shows an excerpt from the motifs decision record. The Chinese strings are treated as evidence objects and interpreted through the scope note and hierarchy.

4.1.4 Governance-Relevant Risks in Non-Preferred Terms

The main concentration of risk in term alignment lay in non-preferred terms. Although the adjudicated non-pre-

Table 2. Excerpt from the term-alignment decision record for motifs (AAT ID: 300009700).

Proposed status	Human baseline label (TC)	AI suggestion (TC)	Provenance	Status upheld in review (1/0)	Review rationale/risk summary
Preferred term	母題 (lit. theme; iconography sense)	紋飾 (lit. ornament)	Wikidata	1	Preferred label chosen to match AAT's Design Elements placement and the "decorative-unit" sense, and to maximize target-locale adoptability (neutral register; common in Taiwan usage).
Non-preferred term	紋飾 (lit. ornament)	文樣 (lit. decorative pattern)	Wikidata	0	Semantically close to the decorative-unit sense, but not adopted as a formal NPT because its Japan-derived register has limited frequency and local adoptability in Taiwan Traditional Chinese academic contexts. It may be retained as an internal regional-variant or cross-linguistic note, not as an access point.
Non-preferred term	紋飾 (lit. ornament)	花紋 (lit. patterning)	Wikidata	0	Overly broad; risks boundary blur between a single decorative unit (motif) and broader pattern configurations (pattern).
Non-preferred term	主題 (lit. topic)	母題 (lit. theme; iconography sense)	NAER Web of Words	1	Accepted as an access point in iconography-related discourse, but not as the core "decorative-unit" definition under the AAT Design Elements context; therefore, retained as non-preferred, not preferred.
Non-preferred term	—	圖案 (lit. design)	Cambridge Dictionary	0	Too broad; confusable with patterns; weak boundary control in thesaurus semantics.
Disfavored term	—	基調 (lit. tonal keynote)	Collins Online Dictionary; NAER Web of Words	1	Cross-domain drift toward a music/color register; excluded to avoid misleading access points.
Disfavored term	—	音樂動機 (lit. musical motif)	Wikidata	1	Domain-specific music term; excluded because it deviates from the visual decorative definition.

Note. TC, Traditional Chinese; AAT, Art & Architecture Thesaurus; AI, artificial intelligence; NPT, non-preferred term; NAER, National Academy for Educational Research. "Status upheld in review" indicates whether the candidate's proposed status in the decision record—preferred, non-preferred, or disfavored—was upheld during editorial review under the scope-note and hierarchy constraints. A value of 1 does not necessarily indicate adoption as the preferred label; for disfavored rows, it indicates that exclusion was upheld.

ferred-term alignment macro-average was 71.9% in this sample, adoption into the controlled vocabulary was more selective after editorial review. This difference shows that linguistic plausibility alone is not sufficient for thesaurus governance. Adoption also depends on boundary control, synonym policy, and local implementability. Initial disagreement was concentrated in 10 of 106 candidate-level term judgments (9.4%). Eight of these cases were adjudicated as non-adoptable, suggesting that pre-consensus disagreement can function as a practical review cue in synonym governance.

The motifs case illustrates two recurrent risk types. The first is semantic over-breadth. Everyday terms such as 花紋 (patterning) and 圖案 (design) can refer both to a single decorative unit and to a broader pattern configuration, whereas AAT distinguishes motifs from patterns. The second is constrained local implementability. 文樣 (decorative pattern) is semantically close, but its Japan-derived register is less common in Taiwan academic use and may be better treated as a regional variant note than as a core access point. More broadly, the disagreement cases clustered around sense disambiguation, historically or regionally marked variants, and hierarchy-dependent labels with incomplete contextual cues.

Accordingly, the review process separates recommended non-preferred terms from disfavored candidates. Candidates that preserve the concept boundary and support retrieval can be retained as access points, whereas candidates showing over-breadth, cross-domain drift, or low local implementability are flagged for closer editor verification. In this study, governance of non-preferred terms served as a sensitive test of whether AI-generated candidate lists could remain traceable and reviewable under high-volume suggestion conditions.

4.2 Scope Note Translation Outcomes

4.2.1 Exploratory MQM-Based Comparison Against Existing Translations

Using the 30 paired items and author-editor consensus scores, paired-sample t tests summarize observed differences within the present review setting. Across the three MQM-derived dimensions, AI-drafted scope notes received higher adjudicated consensus ratings than the existing translations. The observed mean difference for semantic accuracy was 0.73, $t(29) = 4.25$, $p < 0.001$, $d = 0.78$. Fluency showed an observed mean difference of 0.87, $t(29) = 6.50$, $p < 0.001$, $d = 1.19$, but the zero variance in AI fluency scores (standard deviation [SD] = 0.00) points to a possible ceiling effect under the present rubric and consensus procedure. Cultural adaptation showed an observed mean difference of 0.74, $t(29) = 5.12$, $p < 0.001$, $d = 0.93$. Table 3 reports these exploratory paired-sample summaries.

The zero variance in AI fluency scores points to a possible ceiling effect under the present rubric, the limited sensitivity of the fluency scale at the high-quality end, and the tendency of consensus adjudication to reduce rating variability. The contrast between uniformly high fluency ratings and the remaining concerns about semantic and cultural dimensions suggests that readability is not the same as boundary safety. Within this review setting, the observed differences indicate that AI-drafted scope notes may provide reviewable starting texts.

4.2.2 Communicative Translation as an Operational Drafting Constraint

The scope note task adopted a communicative translation orientation that emphasized readability and appropriate definitional register while preserving conceptual boundaries. Within this drafting constraint, AI-drafted scope notes were rated higher on all three dimensions. In editorial terms, the drafts may provide a reviewable first version that allows editors to focus on boundary-sensitive revisions, cross-entry consistency, and culture-loaded phrasing, while leaving final editorial judgment in place.

4.2.3 Qualitative Strengths Observed in AI-Generated Scope Notes

Qualitative review identified three recurring strengths in the AI drafts within this sample: stable definitional register, lexical choices that reduced ambiguity, and pragmatic clarification of implicit functional constraints. These observations help explain why some AI drafts were easier to review and are considered together with the risk patterns reported below. Examples include Ming (AAT ID 300018438), tilings (AAT ID 300265619), and chapels of ease (AAT ID 300007452).

4.3 Observed Risk Patterns in Scope Note Translation

Qualitative analysis identified two recurrent risk patterns in information-density management within the sampled scope notes: semantic compression and pragmatic over-extension. Both patterns may improve readability or definitional consistency in some cases, but both can also shift scope boundaries and therefore require editorial review.

Semantic compression condenses concrete statements into more abstract phrasing. In some cases, this is acceptable because it supports a more consistent definitional register. For Marxism (AAT ID 300055528), the AI recast a source phrase into a more abstract form that aligned with parallel theoretical terms in the entry. However, compression becomes high risk when it removes details needed for scope control. In silhouette animation (AAT ID 300263870), the description of single-frame exposures was reduced to a generic statement about frame-by-frame shooting, weakening the craft-specific boundary.

Pragmatic over-extension adds plausible background, function, or domain knowledge that is not stated in the source scope note. Such additions should be removed unless they can be back-linked to authoritative evidence, as in amphorae (AAT ID 300148696).

Table 3. Scope note translation ratings (exploratory paired-sample summaries, N = 30).

Dimension	AI M (SD)	Human M (SD)	ΔM (AI − human)	t(29)	*p*	d_z
Semantic accuracy	4.83 (0.38)	4.10 (0.77)	0.73	4.25	<0.001	0.78
Fluency	5.00 (0.00)	4.13 (0.45)	0.87	6.50	<0.001	1.19
Cultural adaptation	4.97 (0.18)	4.23 (0.43)	0.74	5.12	<0.001	0.93

Note. N = 30 paired items. *p* is two-tailed. ΔM = AI mean – human mean. Effect size is Cohen's d_z; M = mean; SD = standard deviation. The tests are exploratory summaries based on author-editor adjudicated consensus scores, not independent external ratings.

These two patterns were treated as governance-relevant review cues instead of automatic error labels; their significance depends on whether the added, omitted, or compressed information affects the AAT concept boundary.

4.4 Summary Aligned to Research Questions

For RQ1, preferred-term alignment was stable in the 30 sampled concepts under the controlled evidence setting. Non-preferred-term alignment was more variable, with an adjudicated macro-average of 71.9%. Evidence-backed candidates showed higher observed alignment than model-inferred candidates (83.7% vs. 61.4%). Within this sample, this provenance-related difference helps prioritize editorial attention.

For RQ2, AI-drafted scope note translations received higher author-editor adjudicated consensus ratings than the existing translations on semantic accuracy, fluency, and cultural adaptation in this 30-item sample. The result supports the use of AI drafts as reviewable starting texts under the present design, especially when semantic and cultural boundary checks remain part of adjudication.

For RQ3, process logs suggest bounded review-stage workload redistribution after evidence preparation and AI drafting. The comparison of about 150 minutes per concept in the traditional process and about 10 minutes in review and finalization excludes evidence package construction, prompt preparation, source selection, and longer-term maintenance. It therefore points to possible redistribution of editorial attention during review and finalization, rather than full-process cost.

5. Discussion

5.1 SimRAG and Warrant-Based Traceable Governance

This study contributes chiefly by recasting AI-assisted multilingual thesaurus localization as warrant-based traceable governance. In thesaurus localization, warrant is enacted through editorial judgment: deciding which evidence is relevant, how a candidate expression fits the AAT concept boundary and hierarchy, whether it is usable in the target-language community, and what KOS status it should receive (Beghtol, 2002; Bullard, 2017; Hjørland, 2002; Kwaśnik, 2010). This judgment is not only procedural. It operates within a structured KOS, where concepts, terms, warrants, and domain-based knowledge units are represented, related, and revised over time (Smiraglia, 2014, 2016).

The key governed unit in SimRAG is therefore the candidate-to-status decision. A model suggestion becomes meaningful for thesaurus governance only when it is reviewed as a candidate and assigned a status: preferred term, non-preferred access point, contextual variant, or rejected candidate. This status assignment matters because preferred, non-preferred, and rejected terms shape different paths of retrieval and interpretation. In this sense, SimRAG supports accountable representation by making the grounds of naming and classification more visible (Mai, 2010; Olson, 2002). The decision record links the evidence source, candidate provenance, risk rationale, editorial review, and final status, so that later editors can see the selected label, the alternatives considered, and the reasons for retaining, downgrading, annotating, or rejecting them.

Seen in this way, the empirical findings support the design claim: a traceable review design can organize editorial attention around evidence, risk, and status assignment. The observed stability of preferred-term alignment, the provenance-related differences among non-preferred candidates, and the higher adjudicated ratings for scope note drafts show where such a design made decisions inspectable in this sample. A useful lesson is that the wider synonym space carried more editorial risk than the preferred-label decision. Ambiguity remains in this process. SimRAG gives editors a way to see where that ambiguity enters the record and how it was handled.

Because KOS decisions remain part of an evolving scheme history, such records also respond to the problem of subject ontogeny and collocative integrity in long-lived schemes: later editors need evidence of how labels, statuses, and rationales changed in order to preserve retrieval function over time (Tennis, 2007, 2012). Smiraglia's account of knowledge evolution across KOSs further supports this point: a KOS can be studied by examining how warranted concepts, terms, and domain-based knowledge units are represented and reconfigured across time (Smiraglia, 2016). In the multilingual thesaurus setting, SimRAG extends this insight from scheme-level versioning and knowledge evolution to candidate-level decision records. Each proposed label or scope-note draft is recorded not merely as an AI output, but as a structured decision object whose

Table 4. Five preferred-term governance patterns derived from the 30 sampled concepts.

Pattern	Decision basis/output	Concept label	Preferred label (Traditional Chinese, human → AI)	Governance implication
Institutional register alignment	Local regulations or official institutional usage; Prioritize official wording	cultural heritage	Human PT: 文化遺產 (lit. cultural heritage) → AI PT: 文化資產 (lit. cultural assets)	Reduce cross-institution variation and strengthen links to policy context.
Scope-note boundary alignment	Scope note constraints and sense boundaries; Use definitional evidence to narrow boundaries	motifs	Human PT: 母題 (lit. theme) → AI PT: 紋飾 (lit. ornament)	Reduce sense misplacement and improve boundary precision.
Cross-cultural register adaptation	Source-vocabulary context and target-community conventions; Risk splitting (preferred, non-preferred, disfavored)	tea masters	Human PT: 茶藝師 (lit. tea-arts specialist) → AI PT: 茶人 (lit. tea person; gloss: tea practitioner)	Prefer labels consistent with the defined target locale; record alternative variants as non-preferred labels when they support retrieval, and flag disfavored terms to prevent role or sense misplacement.
Hierarchical consistency validation	Broader-term or adjacent-term context; Hierarchy-aware validation	Ink-bamboo	Human PT: 墨竹畫 (lit. ink bamboo painting) → AI PT: 墨竹 (lit. ink bamboo)	Prevent category drift and misplacement across levels.
Gap filling under governance	Missing translations or low-consensus entries; Provide candidates with provenance and risk flags	chapels of ease	Human PT: 小教堂 (lit. small chapel) → AI PT: 分堂 (lit. branch chapel; gloss: parish chapel)	Provide a traceable starting point that may reduce editorial search burden at scale.

Note. Cases illustrate governance decision patterns, not a literal-gloss ranking. Chinese strings are Traditional Chinese preferred labels; glosses indicate intended sense shifts and are not AI-generated English translations. PT, preferred term.

evidence, risk, and status can be revisited as the vocabulary evolves. The practical implication is a more inspectable review process. Editors can focus on candidates and drafts whose evidence basis, boundary fit, or cultural usability is uncertain.

5.2 Governance Patterns for Preferred-Term Decisions

Section 5.1 defined the candidate-to-status decision as the governed unit. The five patterns in Table 4 show how this unit operated for preferred terms. The patterns are used here as recurrent grounds on which editors authorized or declined preferred-term status: institutional register, scope-note boundary, target-locale convention, hierarchical fit, and gap filling under governance. In each pattern, the AI output becomes useful when it makes the candidate rationale and evidence basis more visible for editorial judgment.

5.3 A Risk-Tier Rubric for Non-Preferred-Term Governance

Non-preferred terms provide a sensitive site for KOS governance because they expand access while also shaping the interpretive neighborhood of a concept. Adoption depends on candidate acceptability, boundary control, synonym policy, and target-locale implementability. For this reason, the semantic-pragmatic risk pyramid is used as a set of review categories, not technical confidence scores (Fig. 3).

The semantic dimension asks whether a candidate preserves the scope-note and hierarchy-defined boundaries of the AAT concept. The pragmatic dimension asks whether the candidate is usable in the target community, considering register, frequency, regional restriction, and temporal appropriateness. No-risk candidates may be retained as access points; conditionally acceptable candidates preserve meaning but show lower target-locale adoptability; moderate-risk candidates show over-breadth or restricted usage conditions and require closer review; high-risk candidates show register or domain mismatch; and prohibited candidates are unsuitable for controlled-vocabulary use.

By recording reasons such as semantic over-breadth, cross-domain drift, or constrained local implementability as reusable editorial labels, the rubric turns case-level rationales into reviewable governance evidence. It supports transparency and trust without making the risk tier an automatic decision rule: a candidate may be retained as an access point, downgraded to a contextual variant, annotated as regionally restricted, or rejected according to the documented rationale.

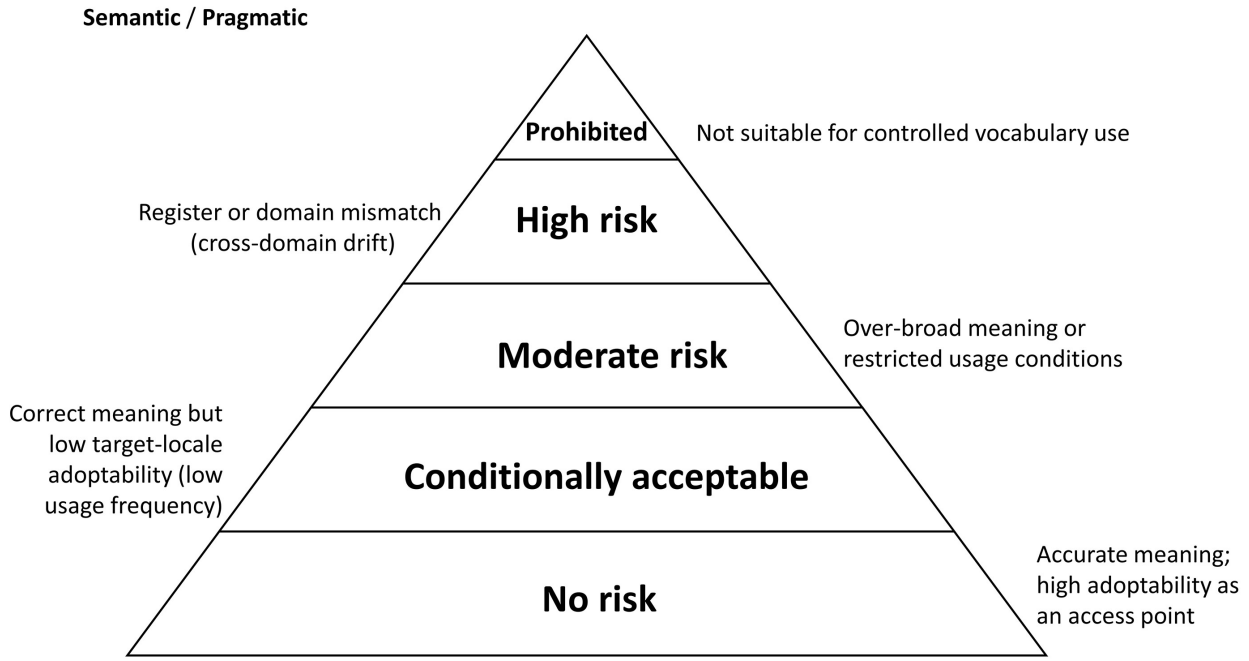


Fig. 3. Semantic-pragmatic risk pyramid for non-preferred-term governance. Note. The pyramid links semantic boundary control with pragmatic suitability. The tiers are review categories for editorial prioritization, not automatic exclusion rules.

5.4 Managing Polysemy and Qualifiers as KOS Boundary Rules

Polysemous labels show that disambiguation is both a linguistic and governance issue. When the same or similar form can denote different concepts across facets, such as an activity and an object, an unqualified expression may be familiar but insufficient as a preferred term. Qualifiers therefore function as KOS boundary rules: they specify which form is authorized as the preferred label and which familiar form may remain as a non-preferred access point. Once prompts explicitly required qualifier-based disambiguation, the model followed the stated editorial rule. This shows that explicit editorial rules can make disambiguation reviewable while leaving policy decisions to editors. Consistent with Hjørland's (2007) view that semantic organization is shaped by classificatory purposes and epistemic perspectives, qualifier rules should be treated as part of the policy layer and carried forward in prompts, review checklists, and later version review.

5.5 Model-Assisted Scope Note Assessment as Diagnostic Feedback

For scope notes, model-assisted pre-assessment functioned as diagnostic feedback before author-editor adjudication. It helped mark possible issues, compare blinded versions, and identify places where semantic accuracy, cultural adaptation, or definitional register required closer review. Final judgments remained author-editor consensus decisions.

The disagreement pattern is therefore used here as a governance signal. No pre-consensus disagreement occurred in fluency, whereas the remaining disagreements concentrated in semantic accuracy and cultural adaptation, especially where technical source terms, semantic compression, or culture-loaded framing required interpretation. This pattern is consistent with the ceiling effects discussed above: fluency may be easier to rate uniformly, while semantic and cultural dimensions remain more boundary-sensitive.

Operational use of model-assisted pre-assessment therefore requires calibration, context provision, and procedural safeguards. Model scoring may favor outputs closer to the model's preferred definitional register, and it may under-detect subtle boundary shifts at the high-quality end. Severity-weighted criteria and periodic human calibration are needed for more routine operational use (Knowles and Lo, 2024; Koo et al., 2024).

5.6 Transferability and Conditions for Reuse in Other KOS Settings

Although the empirical setting is English-to-Traditional-Chinese localization of AAT concepts, the governance logic is not language-specific. It depends on whether a vocabulary program can define reviewable task units, connect candidates to evidence, record provenance and risk, and assign candidates to explicit KOS statuses. The approach may therefore be adapted to other thesauri, controlled vocabularies, or multilingual terminology programs when comparable governance conditions are present.

Table 5. Sample list (index).

AAT ID	Concept label (English)	Facet (AAT code)	Level	Cultural reference
300265422	cultural heritage	Associated Concepts (B.BM)	Upper	Culture-general
300009700	motifs	Physical Attributes (D.DG)	Upper	Culture-general
300264736	Modern	Styles and Periods (F.FL)	Upper	Culture-general
300188613	barbers	Agents (H.HG)	Upper	Culture-general
300056257	decoration	Activities (K.KT)	Upper	Culture-general
300264870	color	Materials (M.MT)	Upper	Culture-general
300033618	paintings	Objects (V.VC)	Upper	Culture-general
300008932	cultural landscapes	Objects (V.RD)	Upper	Culture-general
300014063	textiles	Objects (V.PE)	Upper	Culture-general
300138758	masks	Objects (V.TE)	Upper	Culture-general
300055528	Marxism	Associated Concepts (B.BM)	Lower	Culture-general
300265619	tilings	Physical Attributes (D.DG)	Lower	Culture-general
300264818	Ink-bamboo	Styles and Periods (F.FL)	Lower	East Asia-related
300266039	theater companies	Agents (H.HN)	Lower	Culture-general
300263870	silhouette animation	Activities (K.KT)	Lower	Western-related
300343594	vitreous paint	Materials (M.MT)	Lower	Western-related
300198931	potpourri vases	Objects (V.TQ)	Lower	Western-related
300007452	chapels of ease	Objects (V.RK)	Lower	Western-related
300262837	retablos	Objects (V.VC)	Lower	Western-related
300179434	portfolios	Objects (V.PC)	Lower	Western-related
300018438	Ming	Styles and Periods (F.FL)	Lower	East Asia-related
300106504	Shanghai school	Styles and Periods (F.FL)	Lower	East Asia-related
300073738	Buddhism	Associated Concepts (B.BM)	Upper	East Asia-related
300264376	bugaku	Activities (K.KD)	Upper	East Asia-related
300310956	magatama	Objects (V.VC)	Lower	East Asia-related
300343833	tea masters	Agents (H.HG)	Lower	East Asia-related
300010045	ankhs	Objects (V.VC)	Lower	Western-related
300148696	amphorae	Objects (V.TQ)	Lower	Western-related
300010665	soft paste porcelain	Materials (M.MT)	Lower	Western-related
300042101	lutes	Objects (V.TT)	Lower	Western-related

Note. This appendix lists the 30 sampled AAT concepts used across tasks.

Five conditions are especially important for reuse. First, the vocabulary should provide stable concept identifiers and scope notes or equivalent definitional texts. Second, editors should be able to identify authoritative sources that can support or challenge candidate terms and scope-note phrasings. Third, the vocabulary program should have explicit editorial rules for preferred terms, non-preferred access points, rejected candidates, and notes. Fourth, reviewers need domain and target-language expertise to judge boundary fit, register, and cultural usability. Fifth, the process should preserve decision records so that later editors can reconstruct why a label or draft was accepted, revised, downgraded, annotated, or rejected.

The approach may be less effective where these conditions are absent. If concepts lack stable definitions, sources are sparse or contested, or editorial authority is unclear, SimRAG may still help organize evidence, but comparable review outcomes may not occur. For non-Chinese readers, the Chinese strings in this study should be treated as evidence objects anchored by AAT concept identifiers, English labels, glosses, and decision rationales. This structure can be reused with other language pairs if evidence links, review notes, and final KOS statuses are made visible in a comparable way (Baca and Gill, 2015).

6. Conclusions

This exploratory study examined AI-assisted multilingual thesaurus localization using the AAT as the empirical setting. It places localization in two connected registers: translation and warrant-based traceable governance. SimRAG was proposed as an interim editorial design before full RAG deployment. By linking curated evidence, model suggestions, risk rationales, and editorial decisions, it makes candidate-to-status decisions more inspectable under controlled evidence conditions.

Three bounded conclusions follow. First, in this 30-concept internally adjudicated sample, preferred-term alignment was stable, while non-preferred-term alignment was more variable. The observed provenance-related difference between evidence-backed and model-inferred can-

didates suggests that provenance can function as a review cue within this sample; no general accuracy estimate is implied. Second, AI-drafted scope notes received higher author-editor adjudicated consensus ratings under the present review setting. These ratings provide exploratory evidence that AI drafts may offer reviewable starting texts in this setting; a broader comparative claim would require a different evaluation design. Third, risk patterns such as semantic compression and pragmatic overextension show why fluent wording is insufficient for KOS maintenance and why editorial review must remain attentive to concept boundaries, evidence, and target-locale usability.

The practical implication is bounded. Editorial judgment remains in place. SimRAG organizes review so that editors can inspect evidence, provenance, risk rationale, and final KOS status together. The workload observations reported here indicate possible redistribution during review and finalization after evidence preparation and AI drafting; the study does not estimate full cost. In this sense, the main contribution is a decision structure for recording why a label was authorized, retained as a variant, annotated, or rejected, and why a scope note draft was accepted, revised, or rejected.

This study has clear limits. It is based on 30 AAT concepts in an English-to-Traditional-Chinese setting, a controlled evidence context, selected model-task pairings, and author-editor consensus, without independent external evaluation. Consensus adjudication may have reduced observed variability, especially at the high-quality end. Future work should test SimRAG with larger and more varied samples, independent reviewers, other language pairs and KOS domains, additional models, and fully deployed retrieval settings. Future work can also formalize SimRAG decision records as structured KOS provenance data, including candidate status, warrant source, review activity, and revision rationale. Such research can examine how retrieval quality, attribution design, candidate provenance, and evidence-linked decision records support long-term multilingual thesaurus maintenance.

Availability of Data and Materials

All data reported in this paper will be shared by the corresponding author upon reasonable request.

Author Contributions

SJC conceived and designed the research study. SJC and JSW performed the experiments and analyzed the data. SJC and JSW wrote the original draft. SJC led the subsequent scholarly revision and final refinement of the manuscript. Both authors contributed to editorial changes in the manuscript. Both authors read and approved the final manuscript. Both authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Acknowledgment

The authors thank the anonymous reviewers for their constructive comments and suggestions, which helped improve the manuscript.

Funding

This research was funded by the National Science and Technology Council (NSTC), grant number NSTC 114-2410-H-001-033-MY2, and supported by the Academia Sinica Center for Digital Cultures (Grant No. ASCDC-1137-007).

Conflicts of Interest

The authors declare no conflicts of interest.

Declaration of AI and AI-Assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used ChatGPT to assist with language polishing and limited organizational editing. In addition, ChatGPT models were used in the empirical procedure reported in the main text, where they served as AI systems under study for term-alignment generation, scope note drafting, and model-assisted pre-assessment tasks. The authors reviewed and edited all AI-assisted outputs as needed. All scholarly interpretation, editorial judgment, evaluation, and final decisions were made by the authors, who take full responsibility for the content of the manuscript.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/KO52642.

Appendix

Appendix A. See Table 5.

Appendix B. Prompt templates (normalized)

Note. Templates are normalized representations of prompts used in the experiments; formatting is simplified for readability without changing task constraints.

B.1 Term alignment (SimRAG controlled context)

Context: A research team localizes AAT from English into Traditional Chinese for a GLAM vocabulary project.

Task: Select one preferred term (PT), recommended non-preferred terms (NPTs), and disfavored terms for each English concept, using scope-note and evidence snippets.

Key rules:

1. Prioritize spreadsheet candidates with back-linkable sources.

2. Any additional terms must be labeled MODEL-PROPOSED with rationale and risk.

3. Use the scope note (and optional hierarchy context) plus each candidate's definition/domain note to judge equivalence.

4. Record provenance for every output (source name or MODEL-PROPOSED).

5. If no candidate fits, output "no suitable preferred term" with a brief explanation.

6. Regional variants may be recorded as NPTs/notes when they support retrieval, but are not default PTs under the defined governance scope.

Input fields: source-language (SL) term; SL scope note; optional hierarchy context; target-language (TL) candidate list with definition/domain note and source.

Required outputs: PT, provenance and evidence-linked rationale; NPTs (reason and provenance); disfavored terms (risk and provenance) or "no disfavored terms".

B.2 Scope note translation (zero-shot, communicative)

Context: A research team localizes AAT scope notes from English into Traditional Chinese.

Task: Translate the scope note into Traditional Chinese.

Requirements: (i) Do not omit or add unsupported information; (ii) produce fluent Traditional Chinese consistent with definitional register conventions (neutral, non-evaluative tone).

Input: term; scope note (English). Output: scope note (Traditional Chinese).

Appendix C. Model-assisted pre-assessment prompt (MQM-derived)

Note. This appendix summarizes the MQM-derived rubric used in model-assisted pre-assessment. A worked example of the procedure is provided in **Supplementary Material**.

Role: Professional translation quality assessor. Perform error marking and dimension-level scoring for two blinded versions (A/B).

Goal: Assess whether the translation meets AAT scope note editorial conventions using three dimensions: Accuracy, Fluency, Cultural Adaptation.

Step 1 (Yes/No error marking). Accuracy: Mistranslation; Omission; Addition; Overtranslation; Undertranslation; Untranslated; Do Not Translate. Fluency: Grammar; Word Order; Consistency; Clarity; Coherence. Cultural Adaptation: Locale Conventions; Cultural Context Equivalence; Contextual Suitability.

Step 2 (1–5 scores). Assign Accuracy, Fluency, Cultural Adaptation scores using anchors: 5 no obvious errors; 4 minor issues; 3 noticeable issues; 2 multiple issues; 1 largely unacceptable.

Step 3 (narrative). Provide (i) a brief overall comment; (ii) explanations of markings with cited source phrases.

References

Aixelá JF. Culture-specific items in translation. In Álvarez R, África Vidal MC (eds.) Translation, power, subversion (pp. 52–78). Multilingual Matters: Bristol, UK. 1999. https://doi.org/10.21832/9781800417915-005

Asai A, Wu Z, Wang Y, Sil A, Hajishirzi H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In International Conference on Learning Representations. 2024.

Baca M, Gill M. Encoding Multilingual Knowledge Systems in the Digital Age: the Getty Vocabularies. KNOWLEDGE ORGANIZATION. 2015; 42: 232–243. https://doi.org/10.5771/0943-7444-2015-4-232

Beghtol C. A proposed ethical warrant for global knowledge representation and organization systems. Journal of Documentation. 2002; 58: 507–532. https://doi.org/10.1108/00220410210441

Beghtol C. Domain analysis, literary warrant, and consensus: The case of fiction studies. Journal of the American Society for Information Science. 1995; 46: 30–44. https://doi.org/10.1002/(sici)1097-4571(199501)46:1<30::aid-asi4>3.0.co;2-f

Budimir B. The challenge of translating culture-specific items: evaluating MT and LLMs compared to human translators. In Bouillon P, et al. (eds.) Proceedings of Machine Translation Summit XX: Volume 1 (pp. 455–467). European Association for Machine Translation. 2025. Available at: https://aclanthology.org/2025.mtsummit-1.36/ (Accessed: 3 April 2026).

Bullard J. Warrant as a means to study classification system design. Journal of Documentation. 2017; 73: 75–90. https://doi.org/10.1108/jd-06-2016-0074

Caracciolo C, Stellato A, Rajbahndari S, Morshed A, Johannsen G, Keizer J, et al. Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case. International Journal of Metadata, Semantics and Ontologies. 2012; 7: 65–75. https://doi.org/10.1504/ijmso.2012.048511

Chen GH, Chen S, Liu Z, Jiang F, Wang B. Humans or LLMs as the judge? A study on judgement bias. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 8301–8327). Association for Computational Linguistics: Stroudsburg, PA, USA. 2024. https://doi.org/10.18653/v1/2024.emnlp-main.474

Chen SJ, Zeng ML, Chen HH. Alignment of conceptual structures in controlled vocabularies in the domain of Chinese art: a discussion of issues and patterns. International Journal on Digital Libraries. 2016; 17: 23–38. https://doi.org/10.1007/s00799-015-0163-1

Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960; 20: 37–46. https://doi.org/10.1177/001316446002000104

Doerr M. The CIDOC Conceptual Reference Module: an ontological approach to semantic interoperability of metadata. AI Magazine. 2003; 24: 75–92. https://doi.org/10.1609/aimag.v24i3.1720

Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971; 76: 378–382. https://doi.org/10.1037/h0031619

Gao L, Dai Z, Pasupat P, Chen A, Chaganty AT, Fan Y, et al. RARR: Researching and revising what language models

say, using language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 16477–16508). Association for Computational Linguistics: Stroudsburg, PA, USA. 2023a. https://doi.org/10.18653/v1/2023.acl-long.910

Gao T, Yen H, Yu J, Chen D. Enabling large language models to generate text with citations. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 6465–6488). Association for Computational Linguistics: Stroudsburg, PA, USA. 2023b. https://doi.org/10.18653/v1/2023.emnlp-main.398

Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: a survey. arXiv. 2024. https://doi.org/10.48550/arXiv.2312.10997 (preprint)

Getty Vocabulary Program. Scope note. In Getty Vocabulary Program editorial guidelines: Art & Architecture Thesaurus (AAT), section 3.4.1. 2024. Available at: https://www.getty.edu/publications/vocabularies-editorial-guidelines/aat-guidelines/3_editorial_rules/3.4/ (Accessed: 3 April 2026).

Harpring P. Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works. Getty Research Institute: Los Angeles, CA, USA. 2010.

Hjørland B. Domain analysis in information science: eleven approaches–traditional as well as innovative. Journal of Documentation. 2002; 58: 422–462. https://doi.org/10.1108/00220410210431136

Hjørland B. Semantics and knowledge organization. Annual Review of Information Science and Technology. 2007; 41: 367–405. https://doi.org/10.1002/aris.2007.1440410115

International Organization for Standardization. Information and documentation—thesauri and interoperability with other vocabularies—Part 1: Thesauri for information retrieval. ISO 25964-1. International Organization for Standardization. 2011. Available at: https://www.iso.org/standard/53657.html (Accessed: 3 April 2026).

Knowles R, Lo CK. Calibration and context in human evaluation of machine translation. Natural Language Processing. 2024; 31: 1017–1041. https://doi.org/10.1017/nlp.2024.5

Koo R, Lee M, Raheja V, Park JI, Kim ZM, Kang D. Benchmarking cognitive biases in large language models as evaluators. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 517–545). Association for Computational Linguistics: Stroudsburg, PA, USA. 2024. https://doi.org/10.18653/v1/2024.findings-acl.29

Kraus F, Blumenröhr N, Tonne D, Streit A. Mind the Language Gap in Digital Humanities: LLM-Aided Translation of SKOS Thesauri. Anthology of Computers and the Humanities. 2025; 3: 882–903. https://doi.org/10.63744/a98mev2x4ugw

Kwaśnik BH. Semantic Warrant: A Pivotal Concept for Our Field. KNOWLEDGE ORGANIZATION. 2010; 37: 106–110. https://doi.org/10.5771/0943-7444-2010-2-106

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems. 2020; 33: 9459–9474.

Lommel A, Popović M, Burchardt A. Assessing inter-annotator agreement for translation error annotation. In Proceedings of the Workshop on Automatic and Manual Metrics for Operational Translation Evaluation (MTE 2014) (pp. 31–37). Reykjavík, Iceland. 2014. Available at: https://mte2014.github.io/MTE2014-Workshop-Proceedings.pdf (Accessed: 3 April 2026).

Lommel AR, Burchardt A, Uszkoreit H. Multidimensional quality metrics: a flexible system for assessing translation quality. In Proceedings of Translating and the Computer 35. 2013. Available at: https://aclanthology.org/2013.tc-1.6/ (Accessed: 3 April 2026).

Mai JE. Classification in a social world: bias and trust. Journal of Documentation. 2010; 66: 627–642. https://doi.org/10.1108/00220411011066763

Mariana V, Cox T, Melby A. The Multidimensional Quality Metrics (MQM) Framework: a new framework for translation quality assessment. The Journal of Specialised Translation. 2015; 23: 137–161. https://doi.org/10.26034/cm.jostrans.2015.343

Martínez-Ávila D, Budd JM. Epistemic warrant for categorizational activities and the development of controlled vocabularies. Journal of Documentation. 2017; 73: 700–715. https://doi.org/10.1108/jd-10-2016-0129

Miles A, Bechhofer S. SKOS Simple Knowledge Organization System Reference. W3C Recommendation. World Wide Web Consortium. 2009. Available at: https://www.w3.org/TR/skos-reference/ (Accessed: 3 April 2026).

Ménard E, Khashman N, Kochkina S, Torres-Moreno JM, Velazquez-Morales P, Zhou F, et al. A Second Life for TI-IARA: From Bilingual to Multilingual! KNOWLEDGE ORGANIZATION. 2016; 43: 22–34. https://doi.org/10.5771/0943-7444-2016-1-22

Molina L, Hurtado Albir A. Translation Techniques Revisited: A Dynamic and Functionalist Approach. Meta. 2002; 47: 498–512. https://doi.org/10.7202/008033ar

Naveen P, Trojovský P. Overview and challenges of machine translation for contextually appropriate translations. iScience. 2024; 27: 110878. https://doi.org/10.1016/j.isci.2024.110878

Olson HA. The power to name: locating the limits of subject representation in libraries. Kluwer Academic Publishers: Dordrecht, the Netherlands. 2002. https://doi.org/10.1007/978-94-017-3435-6

Rashkin H, Nikolaev V, Lamm M, Aroyo L, Collins M, Das D, et al. Measuring Attribution in Natural Language Generation Models. Computational Linguistics. 2023; 49: 777–840. https://doi.org/10.1162/coli_a_00486

Shi L, Ma C, Liang W, Diao X, Ma W, Vosoughi S. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. 2025. https://doi.org/10.18653/v1/2025.

ijcnlp-long.18

Silva B, Buchicchio M, Van Stigt D, Stewart C, Moniz H, Lavie A. Data-driven Asian Adapted MQM Typology and Automation in Translation Quality Workflows. The Journal of Specialised Translation. 2024; 41: 98–126. https://doi.org/10.26034/cm.jostrans.2024.4713

Singh P, Patidar M, Vig L. Translating across cultures: LLMs for intralingual cultural adaptation. In Proceedings of the 28th Conference on Computational Natural Language Learning (pp. 400–418). Association for Computational Linguistics: Stroudsburg, PA, USA. 2024. https://doi.org/10.18653/v1/2024.conll-1.30

Slobodkin A, Hirsch E, Cattan A, Schuster T, Dagan I. Attribute first, then generate: locally-attributable grounded text generation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 3309–3344). Association for Computational Linguistics: Stroudsburg, PA, USA. 2024. https://doi.org/10.18653/v1/2024.acl-long.182

Smiraglia RP. Empirical Methods for Knowledge Evolution across Knowledge Organization Systems. KNOWLEDGE ORGANIZATION. 2016; 43: 351–357. https://doi.org/10.5771/0943-7444-2016-5-351

Smiraglia RP. The elements of knowledge organization. Springer International Publishing: Cham. 2014. https://doi.org/10.1007/978-3-319-09357-4

Stellato A, Rajbhandari S, Turbati A, Fiorelli M, Caracciolo C, Lorenzetti T, et al. VocBench: a web application for collaborative development of multilingual thesauri. In European Semantic Web Conference (pp. 38–53). Springer International Publishing: Cham. 2015. https://doi.org/10.1007/978-3-319-18818-8_3

Tennis JT. Scheme Versioning in the Semantic Web. Cataloging & Classification Quarterly. 2007; 43: 85–104. https://doi.org/10.1300/j104v43n03_05

Tennis JT. The strange case of eugenics: A subject's ontogeny in a long-lived classification scheme and the question of collocative integrity. Journal of the American Society for Information Science and Technology. 2012; 63: 1350–1359. https://doi.org/10.1002/asi.22686

Walhain L, Albouze S, Gerencsér A, Paunescu M, Tzouvaras V, Palma C. The EuroVoc thesaurus: management, applications, and future directions. In Alam M, Tchechmedjiev A, Gracia J, Gromann D, di Buono MP, Monti J, et al. (eds.) Proceedings of the 5th Conference on Language, Data and Knowledge (pp. 340–350). Unior Press: Naples, Italy. 2025. Available at: https://aclanthology.org/2025.ldk-1.34/ (Accessed: 3 April 2026).

Wataoka K, Takahashi T, Ri R. Self-preference bias in LLM-as-a-judge. arXiv. 2024. https://doi.org/10.48550/arXiv.2410.21819 (preprint)

Yan J, Yan P, Chen Y, Li J, Zhu X, Zhang Y. GPT-4 vs. human translators: a comprehensive evaluation of translation quality across languages, domains, and expertise levels. arXiv. 2024. https://doi.org/10.48550/arXiv.2407.03658 (preprint)

Yao B, Jiang M, Bobinac T, Yang D, Hu J. Benchmarking machine translation with cultural awareness. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 13078–13096). Association for Computational Linguistics: Stroudsburg, PA, USA. 2024. https://doi.org/10.18653/v1/2024.findings-emnlp.765

Zeng ML, Chan LM. Trends and issues in establishing interoperability among knowledge organization systems. Journal of the American Society for Information Science and Technology. 2004; 55: 377–395. https://doi.org/10.1002/asi.10387

Zeng ML, Mayr P. Knowledge Organization Systems (KOS) in the Semantic Web: a multi-dimensional review. International Journal on Digital Libraries. 2019; 20: 209–230. https://doi.org/10.1007/s00799-018-0241-2