International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

RESEARCH ON THE CONSTRUCTION OF A KNOWLEDGE BASE FOR NEW CHINESE WORDS BASED ON LARGE LANGUAGE MODELS

Camilla Quaranta *

Henan University, Kaifeng, Henan Province, China.
* Corresponding Author

International Journal of Science and Research Archive, 2026, 21(01), 672–681

Publication history: Received on 19 September 2026; revised on 25 October 2026; accepted on 27 October 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.21.1.1942

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

The continuous emergence of new Chinese words and expressions in digital communication presents important challenges for computational lexicography, natural language processing, and knowledge organization. Conventional lexicographic resources cannot always reflect newly emerging lexical forms, evolving meanings, and changes in usage across digital communities. Although large language models (LLMs) offer substantial capabilities for contextual interpretation and structured information extraction, their application to lexical knowledge-base construction requires mechanisms for evidence verification, temporal representation, and linguistic quality control. This study proposes a provenance-aware framework for constructing a dynamic knowledge base of Chinese neologisms by integrating corpus-based candidate discovery, LLM-assisted lexical analysis, retrieval-grounded verification, and temporal knowledge representation. The framework distinguishes new lexical forms from new meanings of existing words, proper names, and non-lexical expressions. A structured knowledge model is developed to represent lexical forms, senses, formation mechanisms, contextual examples, source evidence, and temporal attributes. Additionally, a reproducible evaluation protocol is specified to assess lexical identification, semantic extraction, factual grounding, knowledge-based consistency, and updating performance. The methodological analysis demonstrates how lexical knowledge can be organized into verifiable, temporally qualified records rather than isolated machine-generated definitions. The proposed approach provides a systematic foundation for developing continuously updated Chinese lexical resources and identifies the empirical validation required before deployment.

Keywords: Chinese Neologisms; Large Language Models; Lexical Knowledge Base; Knowledge Graph; Semantic Change; Information Extraction.

1 INTRODUCTION

Language continually evolves in response to technological innovation, cultural exchange, social transformation, and changing communication practices. The expansion of digital media has accelerated the circulation of newly coined words, abbreviations, semantic extensions, and culturally specific expressions. In contemporary Chinese, these developments are especially visible in online discussions, short-video platforms, digital communities, and technology-related discourse. Emerging expressions often acquire meanings that cannot be understood through the conventional interpretation of their individual Chinese characters. Chinese lexicography has long attempted to document lexical innovation through periodically revised dictionaries and collections of new words. Gao [1], in a review of *A Dictionary*

of Chinese Neologisms (2000–2020), highlights the diversity of Chinese lexical innovation, including loanwords, internet expressions, and new meanings assigned to established lexical forms. Nevertheless, the publication cycle of traditional dictionaries creates a structural limitation when newly emerging terms develop and spread rapidly.

The temporal characteristics of lexical innovation further complicate knowledge organization. Jiang et al. [2] examined the diffusion and decline of Chinese neologisms and demonstrated that their usage trajectories frequently exhibit periods of rapid growth followed by decreasing popularity. Consequently, a lexical resource that merely records whether a word exists cannot fully represent when the expression emerged, how its usage developed, or whether it's meaning subsequently changed. Automatic Chinese word discovery provides an important technical foundation for addressing this problem. Pan et al. [3] proposed TopWORDS-Seg, a Bayesian framework for simultaneous Chinese text segmentation and word discovery in open-domain environments. Other studies have investigated neural architectures and lexical enhancement techniques to improve Chinese word segmentation and sequence labeling [4,5]. Although these approaches contribute to identifying lexical boundaries and unregistered expressions, the construction of a comprehensive neologism knowledge base requires additional procedures for meaning interpretation, contextual verification, source attribution, and continuous updating.

Recent research has directly explored the capabilities and limitations of LLMs for Chinese neologism processing. Zou and Yu [6] examined the understanding of Chinese neologisms by large language models. Wu and Liu [7] subsequently introduced a corpus-based Chinese web neologism detection pipeline combining statistical candidate filtering with linguistically informed LLM classification. Zhao et al. [8] developed CNeo-Bench to evaluate LLM understanding of Chinese neologisms across different word-formation mechanisms. These investigations establish that identifying and interpreting Chinese neologisms remains a nontrivial computational problem.

However, lexical detection and lexical knowledge organization are distinct research objectives. Detecting an emerging expression does not automatically determine its validated definition, grammatical characteristics, semantic relationships, historical evolution, or suitability for inclusion in a reusable knowledge resource. Furthermore, LLM-generated lexical explanations may contain plausible but unverified claims, particularly concerning word origins, early attestations, or rapidly evolving cultural meanings. To address these issues, this study develops a methodological framework for constructing a Chinese neologism knowledge base through the integration of statistical discovery, language-model-assisted analysis, corpus evidence, and explicit knowledge representation. The research addresses four questions:

- **RQ1:** How can emerging Chinese lexical forms and new senses of established words be identified from temporally organized textual corpora?
- **RQ2:** How can LLMs generate structured lexical descriptions while maintaining traceability to authentic linguistic evidence?
- **RQ3:** How should a knowledge base represent the semantic, morphological, contextual, and temporal characteristics of Chinese neologisms?
- **RQ4:** What evaluation procedures are necessary to establish the reliability, coverage, consistency, and maintainability of an LLM-assisted lexical knowledge base?

The principal contribution is a unified methodological design in which lexical discovery, semantic interpretation, provenance, and temporal updating are treated as interdependent components. Unlike a static collection of automatically generated word definitions, the proposed model associates each lexical assertion with supporting evidence and an explicit verification status.

2 LITERATURE REVIEW AND THEORETICAL FOUNDATIONS

2.1 Chinese neologisms and lexical evolution

Neologisms may arise through the creation of new lexical forms, the reinterpretation of established words, abbreviation, borrowing, semantic extension, or culturally specific word-formation processes. Chinese neologisms also exhibit characteristics associated with the Chinese writing system, including homophonic substitutions, character combinations, numerical expressions, and Pinyin-based abbreviations. A significant distinction must therefore be maintained between formal novelty and semantic novelty. Formal novelty occurs when a previously unregistered lexical form enters observable language use. Semantic novelty occurs when an existing word acquires an additional meaning or develops a distinct usage convention. For instance, a familiar Chinese expression may gradually acquire a specialized internet meaning while retaining its earlier meanings. A detection system that relies exclusively on whether an expression appears in an established dictionary may fail to identify this semantic development. Hamilton et al. [9] demonstrated the value of diachronic word representations for investigating systematic patterns of semantic change. Schlechtweg et al. [10] further established a benchmark-oriented framework for evaluating lexical semantic change

detection. Although these investigations were not specifically focused on Chinese internet neologisms, their methodological contributions support the analysis of temporally evolving lexical meanings. The implications for knowledge-based construction are substantial. A knowledge resource should represent lexical forms and lexical senses separately, permitting several meanings to coexist under the same written expression. Each sense should also be associated with contextual examples and relevant time intervals.

2.2 Knowledge graphs and structured lexical representation

Knowledge bases transform scattered information into structured representations that can be retrieved, connected, updated, and reused. Knowledge graphs provide a flexible architecture for expressing entities, concepts, attributes, and their semantic relationships. Hogan et al. [11] describe knowledge graphs as frameworks for organizing heterogeneous information through explicit graph-based representations. Vrandečić and Krötzsch [12] illustrate the potential of collaboratively maintained structured knowledge through Wikidata. These approaches are relevant to linguistic knowledge organization because lexical entries involve multiple interconnected information types. A neologism may possess alternative written forms, several senses, associated cultural topics, grammatical functions, and relationships with other lexical expressions.

Recent research has also examined the interaction between LLMs and knowledge-graph construction. Bian [13] surveys LLM-supported ontology development, information extraction, and knowledge fusion. Chen et al. [14] investigate the effectiveness of LLMs as knowledge-graph constructors and highlight the need to assess extracted information from both structural and semantic perspectives. For Chinese neologisms, a knowledge graph can support not only lexical lookup but also queries involving morphological patterns, semantic associations, cultural domains, and historical development. Nevertheless, general-purpose knowledge-graph extraction does not automatically resolve the linguistic distinction between a word, a phrase, and a proper name. This distinction requires domain-specific annotation rules and human linguistic judgment.

2.3 LLM-based information extraction

The transition from manually engineered extraction rules to pretrained and instruction-driven language models has expanded the range of information that can be recovered from unstructured text. Lu et al. [15] developed a unified structure-generation framework for information extraction, illustrating how different extraction objectives can be expressed through a common structured-output architecture. Huguet Cabot and Navigli [16] formulated relation extraction as end-to-end language generation. Wang et al. [17] subsequently investigated instruction tuning for unified information extraction. These methods suggest that lexical attributes can be extracted through task-specific instructions without requiring a separate predictive architecture for every information category. A language model can be instructed to classify lexical novelty, infer contextual meaning, identify grammatical characteristics, and generate candidate semantic relationships. However, a generated answer should not automatically be regarded as verified lexical knowledge. An LLM may infer a convincing definition without identifying whether it is supported by the supplied text. This problem is especially important for recent Chinese expressions whose meanings depend on rapidly changing cultural contexts. Retrieval-augmented generation offers a potential response to this limitation. Lewis et al. [18] introduced an architecture that combines parametric language generation with external knowledge retrieval. Asai et al. [19] further investigated retrieval and self-evaluation mechanisms intended to improve generated responses. In the present framework, retrieved corpus evidence is used not merely to provide background context but also to determine whether individual lexical claims are eligible for storage.

2.4 Research gap

The literature demonstrates substantial progress in Chinese word discovery, LLM-based neologism recognition, semantic change detection, and knowledge-graph construction. Yet these research directions address different stages of the broader lexical knowledge-management problem. Three limitations motivate the proposed framework. First, candidate detection does not necessarily yield a reusable, sense-disambiguated lexical record. Second, generative definitions may lack transparent evidence or introduce unsupported historical claims. Third, resources that do not represent time explicitly cannot adequately distinguish emerging, established, and declining lexical usages. Accordingly, the present study focuses on the integration of form–sense separation, claim-level provenance, temporal qualification, and verification-controlled updates within a single knowledge-based architecture. This contribution is methodological: the proposed integration must be implemented and evaluated before any claim of superior empirical performance can be established.

3 MATERIALS AND METHODS

3.1 Research design

This study adopts a design-oriented methodological approach to specify the construction, validation, and evaluation of an LLM-assisted lexical knowledge base. The proposed system contains five functional layers: corpus acquisition, lexical candidate discovery, LLM-assisted semantic analysis, evidence-based quality control, and temporal knowledge-based storage.

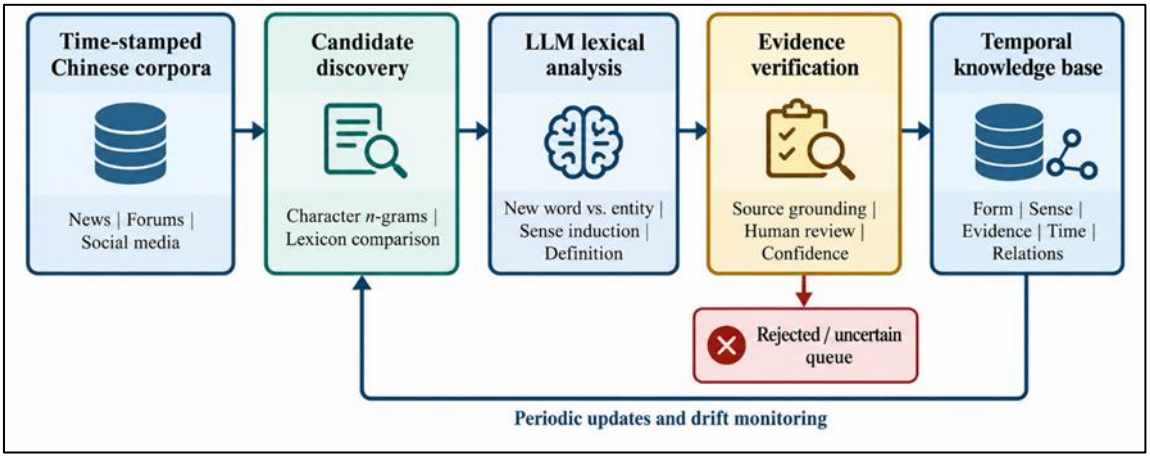


Figure 1: Proposed provenance-aware framework for a Chinese neologism knowledge base.

The architecture is designed to make each knowledge-construction step auditable. Every accepted lexical entry should be traceable to identifiable source documents, recorded extraction procedures, and a verification decision.

3.2 Data acquisition and corpus preparation

The proposed corpus should integrate appropriately licensed Chinese textual resources from several complementary domains, including news articles, public online discussions, cultural publications, technology-related texts, and available lexical databases. Each source document should contain, where available, a unique document identifier, collection date, publication date, source category, text content, language variety, and usage permissions. The corpus should be organized into fixed temporal intervals, such as monthly partitions, to support the observation of lexical emergence and changing frequency. The exact observation window must be defined before empirical analysis and should include both historical and recent material. Text preprocessing should include character-encoding normalization, removal of exact duplicates, identification of near-duplicate documents, and separation of textual material from irrelevant markup. Crucially, normalization should not destroy meaningful differences between character forms, homophonic spellings, numbers, Latin-letter abbreviations, and expressive variants. The original textual form should therefore be preserved alongside a normalized representation. Because online sources may contain personal information, corpus collection should comply with platform access conditions, copyright restrictions, and applicable privacy requirements. Personally, identifying information that is unnecessary for lexical research should not be retained.

3.3 Discovery of candidate neologisms

Chinese word segmentation is challenging because written Chinese generally lacks explicit spaces between lexical units. Pretrained language representations, such as BERT, provide contextual modeling capabilities that can support lexical processing [20]. Sentence-level semantic representations may also facilitate the comparison of contextual examples and the grouping of potentially related lexical senses [21]. Nevertheless, relying exclusively on a pretrained tokenizer would introduce a practical limitation: an emerging expression may be divided into already-known characters or substrings before it can be considered as a lexical candidate. The proposed framework therefore combines tokenizer-independent character-sequence generation with lexical comparison and contextual analysis. Candidate character sequences are extracted from the cleaned corpus, including Chinese-character sequences and mixed-script expressions. Statistical measures such as pointwise mutual information and boundary entropy can then be used to assess the cohesion and independence of candidate expressions. For a candidate sequence w divided into adjacent components a and b, a cohesion measure may be expressed as:

$$\mathbf{PMI}(\boldsymbol{a},\boldsymbol{b}) = \log\left[\frac{\boldsymbol{P}(\boldsymbol{ab})}{\boldsymbol{P}(\boldsymbol{a})\boldsymbol{P}(\boldsymbol{b})}\right]$$

For multi-character candidates, the minimum PMI over permissible internal splits can be used to identify weakly associated character combinations. Boundary entropy is used to examine the diversity of characters appearing immediately before or after the candidate:

$$\boldsymbol{H(X) = -\sum_{x \in X} p\,(x) \log p\,(x)}$$

Higher contextual boundary diversity may support the interpretation of a candidate as an independent lexical unit, although neither PMI nor entropy alone establishes lexical status. Candidate terms should also be compared with selected dictionary inventories. Terms absent from the reference inventory become potential new-form candidates, while dictionary-listed terms should remain eligible for subsequent semantic-novelty analysis. To capture temporal change, candidate frequencies can be calculated for each observation interval. A normalized change score can be defined as:

$$\boldsymbol{B(w,t) = \log\left[\frac{r(w,t) + \varepsilon}{r(w, \text{baseline}) + \varepsilon}\right]}$$

where *r (w, t)* is the normalized occurrence rate of candidate *w* during period *t*, *r (w, baseline)* represents its earlier usage rate, and *ε* is a smoothing constant. Temporal increase provides evidence of changing usage but does not, by itself, establish a new word or a new meaning. All thresholds for candidate generation should be selected on development data rather than the final evaluation set.

3.4 LLM-assisted lexical interpretation

After candidate discovery, an instruction-following LLM should evaluate each candidate using authentic, retrieved contextual examples. Instead of asking the model to produce a general explanation, the framework specifies a constrained extraction task in which the model must return predefined lexical attributes. The required output includes the following information:

- Candidate lexical form and normalized variant.
- Classification as a new form, new sense, established usage, proper name, non-lexical phrase, or uncertain case.
- Context-supported definition.
- Word-formation mechanism.
- Grammatical category, when determinable.
- Typical usage context and pragmatic function.
- Potentially related words or senses.
- Identifiers of evidence supporting each assertion.
- Uncertainty indicators and reasons for non-verification.

The model should be explicitly instructed not to invent word origins, historical emergence dates, named sources, or usage examples. A two-stage semantic interpretation process is recommended. The first stage identifies whether the candidate functions as a lexical unit and distinguishes it from ordinary compositional phrases and named entities. The second stage analyzes its contextual meaning and relationship to previously documented lexical senses. Where an existing word has acquired a novel meaning, the model should propose a new sense record rather than creating an unrelated lexical entry. The output should be validated against a strict machine-readable schema. Responses containing unsupported relation types, absent evidence identifiers, or invalid data types should be rejected or returned for correction.

3.5 Design of the lexical knowledge model

The proposed knowledge base is defined conceptually as a temporally qualified graph:

$$\boldsymbol{G = (V, E, A, T)}$$

where *V* represents nodes, *E* denotes semantic and structural relationships, *A* represents descriptive attributes and evidence metadata, and *T* represents temporal information. The principal node classes are Lexeme, SurfaceForm, Sense, Evidence, Source, and Topic. A Lexeme represents a lexical identity. A SurfaceForm represents an observable orthographic, phonetic, numerical, or abbreviated realization. A Sense represents a distinct contextual meaning. Evidence nodes reference attested usages in specific documents. Source nodes retain information about the original publication or dataset. Representative relationships include: hasForm: connects a lexeme with its written or abbreviated representation. hasSense: associates a lexeme with a particular meaning. Supported By: links a lexical assertion to an authentic evidence record. relatedTo: identifies a verified semantic or topical relationship. derivedFrom:

records a supported lexical formation or source-form relationship. associatedWithTopic: connects an expression to a cultural or thematic domain. supersedes: identifies a later version of a corrected or updated lexical assertion. The graph design distinguishes an assertion from the evidence supporting it. For example, a relationship between a lexical form and a specific sense should possess its own verification status and supporting evidence, rather than relying exclusively on the existence of an entry-level source. This approach is compatible with the principles of provenance representation formalized in the W3C PROV Ontology [22] and with the broader requirements for findable, accessible, interoperable, and reusable data identified by Wilkinson et al. [23]. The knowledge-base model should additionally distinguish three temporal attributes:

- **Observation time:** the date on which a lexical usage was observed in a source.
- **Validity interval:** the period during which a particular usage or sense is supported by the available corpus evidence.
- **Record time:** the date on which the knowledge-based assertion was created, reviewed, or updated.

These attributes prevent a common interpretive error: treating the earliest observation within a sampled corpus as the actual historical date when a word was invented.

3.6 Evidence verification and quality control

The reliability of the proposed knowledge base depends on separating model-generated hypotheses from validated lexical knowledge. Each proposed entry should undergo corpus-grounded verification. Definitions, word-formation labels, contextual interpretations, and semantic relationships should be checked against retrieved textual evidence. Where available, corroborating evidence from independent source origins should be preferred. Repeated copies of the same article should not be treated as independent attestations. The verification mechanism should assign one of three statuses:

- **Verified:** sufficient contextual evidence supports the proposed lexical interpretation.
- **Provisional:** evidence suggests a plausible lexical interpretation, but confirmation is incomplete.
- **Rejected:** the candidate is unsupported, incorrectly classified, or inconsistent with the evidence.

The verification process should include automated structural checks and expert linguistic review. Cases involving polysemy, culturally specific metaphors, ambiguous morphological boundaries, or disputed novelty should receive priority for human adjudication. A numerical confidence score may be estimated from calibrated model probabilities and evidence-based features, but it must not be presented as an objective measure of truth without empirical calibration. Long-context LLM behavior also requires attention. Liu et al. [24] demonstrated that relevant information may be used inconsistently depending on its position within lengthy model inputs. For lexical verification, this suggests the value of retrieving concise, directly relevant evidence passages rather than supplying large, undifferentiated document collections.

3.7 Knowledge-based updating

A dynamic knowledge base should support periodic ingestion of new documents and re-evaluation of previously identified lexical items. The updating mechanism should identify newly emerging candidates, additional attestations of existing entries, potential changes in meaning, and expressions whose usage has decreased. An entry should not be deleted merely because its frequency declines. Lexical disappearance from a particular platform or period does not necessarily establish disappearance from the language. Instead, the framework should preserve historical versions while recording changes in observed usage and interpretation. When an earlier meaning requires revision, the previous verified assertion should remain traceable through version history. Continuous updating should be treated as a controlled evidence-integration process rather than unrestricted automatic rewriting by an LLM.

3.8 Evaluation design

A complete empirical assessment should examine lexical detection, semantic interpretation, knowledge-based quality, temporal reliability, and operational efficiency. The proposed evaluation uses a manually adjudicated reference dataset containing multiple classes of lexical phenomena, including new forms, novel senses, conventional words, named entities, ordinary phrases, and uncertain cases. The reference material should cover multiple periods and source domains to assess generalization. To reduce data leakage, training, development, and testing partitions should be separated at the source-document level, with additional temporal separation for testing newly emerging usages. Three expert annotators are recommended for independently evaluating a stratified sample, followed by adjudication of disagreements. Inter-annotator agreement should be reported using an appropriate reliability statistic, following established guidance on linguistic annotation evaluation [25]. The framework should be compared with methods of increasing complexity: dictionary-based matching, statistical candidate identification, an LLM-only approach, and the complete evidence-grounded framework. Ablation studies should separately remove temporal features, evidence

retrieval, source-level verification, sense separation, and expert review to determine the contribution of each component.

4 ANALYTICAL DEMONSTRATION AND DISCUSSION

4.1 Representation of different types of Chinese neologisms

To illustrate the proposed method, Table 1 presents example lexical cases. These examples describe how different linguistic phenomena should be represented; they are not claimed as records extracted or verified by an implemented system.

Table 1: Illustrative treatment of Chinese lexical phenomena in the proposed knowledge base

Expression	Linguistic phenomenon	Proposed treatment
yyds (永远的神)	Pinyin-initial abbreviation	Connect abbreviated form with expanded expression and supported contextual sense
内卷	Semantic extension	Maintain separate literal/earlier and contemporary sociocultural sense records
躺平	Context-dependent figurative usage	Associate relevant senses with evidence and domain labels
村超	Name associated with a specific event	Route to named-entity classification unless separate lexical usage is established
886	Numerical homophonic expression	Link numerical surface form to a supported spoken expression

The example involving *内卷* demonstrates why a one-word–one-definition structure is insufficient. The same written form can participate in semantically different usages that must not be merged indiscriminately. Likewise, *yyds* illustrates the importance of representing abbreviated forms independently from their expansions while preserving their relationship. The case of *村超* highlights the necessity of distinguishing a named event from a general lexical innovation. Its appearance in a corpus does not automatically establish that it should be classified as a common neologism.

4.2 Comparative analysis of construction strategies

Table 2: Methodological comparison of lexical knowledge-construction strategies

Criterion	**Dictionary matching**	**Statistical discovery**	**LLM-only extraction**	**Proposed framework**
Unregistered-form detection	Limited by dictionary coverage	Supported	Supported	Supported
New-sense representation	Requires additional analysis	Limited	Possible	Explicit sense modeling
Contextual definition	Usually, unavailable	Unavailable	Generated	Evidence-grounded generation
Source provenance	Not intrinsic	Depends on implementation	Not guaranteed	Required for accepted assertions
Temporal representation	Not intrinsic	Possible	Not guaranteed	Explicit
Expert review mechanism	External	External	Optional	Integrated

Traditional dictionary matching offers relatively straightforward candidate filtering, but dictionary inventories alone cannot establish the emergence of new senses. Statistical approaches provide evidence about lexical cohesion and

frequency change, yet these measures do not directly represent cultural meaning or pragmatic usage. LLM-only methods can generate rich lexical descriptions, but model-generated explanations may be unsupported or inconsistent across contexts. The proposed framework attempts to combine these complementary capabilities while making verification an explicit condition of knowledge-based inclusion.

The table describes design capabilities rather than measured performance. It should not be interpreted as evidence that the proposed framework outperforms existing methods.

4.3 Anticipated benefits and failure conditions

The framework provides a systematic mechanism for converting unstructured lexical observations into structured, verifiable knowledge records. Its expected value lies in four areas: improved traceability of lexical definitions, clearer separation between forms and meanings, preservation of historical changes, and more controlled integration of LLM-generated linguistic interpretations. However, several failure conditions remain important. First, frequency-based filtering may underrepresent legitimate but rare expressions. Second, uneven source coverage may distort judgments of lexical popularity. Third, LLMs may fail to interpret culturally specific expressions without sufficient context. Fourth, identical surface forms may carry different meanings across communities. A further challenge concerns feedback effects. If the knowledge base is repeatedly used to provide model context, unverified errors may be reinforced during subsequent extraction cycles. For this reason, provisional and rejected claims should not be silently treated as validated evidence. Automated evaluation alone is insufficient. Research on LLM-based text evaluation, including G-Eval [26], indicates both the potential utility of model-assisted assessment and the importance of accounting for evaluator bias. Similarly, text-overlap metrics such as ROUGE [27] may be useful as supplementary indicators for generated definitions, but they cannot independently establish semantic correctness, cultural appropriateness, or evidential support.

4.4 Evaluation indicators

The following indicators are proposed for a subsequent implementation study.

Table 3: Recommended evaluation dimensions and measurements

Dimension	Measurement	Evaluation objective
Neologism classification	Precision, recall, macro-F1	Correct differentiation of lexical classes
New-sense detection	Sense-level precision, recall, F1	Identification of semantic innovation
Definition quality	Expert semantic accuracy ratings	Faithfulness and contextual appropriateness
Evidence grounding	Supported-claim proportion	Reduction of unsupported lexical assertions
Knowledge-base integrity	Schema violations, contradictory claims	Structural and logical consistency
Temporal accuracy	Correct temporal ordering and interval assessment	Reliability of lexical histories
Coverage	Recall across source domains and formation types	Representativeness
Efficiency	Processing time and computational cost	Operational feasibility

For classification, precision and recall should be calculated using independently adjudicated reference annotations. Macro-F1 is appropriate when different lexical classes exhibit unequal frequencies. Evidence grounding should be evaluated at the level of individual claims. A definition containing several propositions should not be considered fully supported when only one proposition is grounded in an authentic source. The final empirical study should also report uncertainty intervals and differences between baseline systems rather than relying solely on point estimates.

Limitations

The present study specifies a methodological framework and does not report implementation-based performance. Its contribution is therefore restricted to architectural design, formal knowledge organization, and the proposed evaluation procedure. Several practical limitations must be addressed before deployment. The availability and representativeness

of Chinese-language corpora may vary across platforms and regions. Cultural expressions may also be difficult to interpret when relevant social context is missing. Commercial and open-source LLMs can differ substantially in their language competence, model updates, reproducibility, and operational cost. Consequently, findings obtained with one model should not automatically be generalized to other models. Finally, lexical novelty is not a purely computational property. Expert linguistic interpretation remains necessary for deciding whether an observed expression functions as a lexical unit, a new sense, a conventional phrase, or a temporary linguistic innovation.

5 CONCLUSION

This study develops a methodological framework for constructing a dynamic knowledge base of Chinese neologisms through the integration of large language models, corpus-based lexical discovery, contextual verification, and temporal knowledge representation. The proposed architecture addresses several problems that arise when emerging lexical expressions are treated as isolated dictionary entries. It separates lexical forms from their senses, attaches provenance to individual assertions, distinguishes observation dates from historical origins, and provides verification-controlled mechanisms for maintaining evolving lexical records. The analysis establishes a reproducible design for organizing emerging Chinese lexical knowledge but does not establish empirical performance superiority. Future research should implement the architecture, construct a manually validated evaluation dataset, and investigate its effectiveness across different language models, domains, and time periods. With appropriate implementation and validation, this approach may contribute to computational lexicography, language learning applications, semantic retrieval, and the broader development of updatable Chinese-language knowledge resources.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

The authors acknowledge the academic and institutional support provided during the preparation of this manuscript. No additional technical or professional assistance requiring specific acknowledgment was received.

Disclosure of conflict of interest

The authors declare that they have no known financial or non-financial conflicts of interest that could have influenced the research, preparation, or publication of this manuscript.

REFERENCES

[1] Gao Y. Hou Min. A Dictionary of Chinese Neologisms (2000–2020). *Lexikos*. 2024;34(1):189–197. https://doi.org/10.5788/34-1-1904

[2] Jiang M, Shen XY, Ahrens K, Huang CR. Neologisms are epidemic: Modeling the life cycle of neologisms in China 2008–2016. *PLoS One*. 2021;16(2). https://doi.org/10.1371/journal.pone.0245984

[3] Pan C, Sun M, Deng K. TopWORDS-Seg: Simultaneous text segmentation and word discovery for open-domain Chinese texts via Bayesian inference. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 2022. p. 158–169. https://doi.org/10.18653/v1/2022.acl-long.13

[4] Huang K, Yu H, Liu J, Liu W, Cao J, Huang D. Lexicon-based graph convolutional network for Chinese word segmentation. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021. p. 2908–2917. https://doi.org/10.18653/v1/2021.findings-emnlp.248

[5] Liu W, Fu X, Zhang Y, Xiao W. Lexicon-enhanced Chinese sequence labeling using BERT adapter. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 2021. p. 5847–5858. https://doi.org/10.18653/v1/2021.acl-long.454

[6] Zou Y, Yu S. Can large language models understand Chinese neologisms? In: *2025 International Conference on Asian Language Processing (IALP)*. 2025. p. 13–18. https://doi.org/10.1109/IALP68296.2024.11156705

[7] Wu Y, Liu M. Operationalizing linguistic methods through prompt-engineering skills: An automatic Chinese web neologism detection pipeline. *arXiv* [Preprint]. 2026. arXiv:2606.08715. https://arxiv.org/abs/2606.08715

[8] Zhao K, Miao Z, Xie Z, Cao S, Tsuruoka Y. CNeo-Bench: Diagnosing large language models on Chinese neologisms. *arXiv* [Preprint]. 2026. arXiv:2608.28053. https://arxiv.org/abs/2608.28053

[9] Hamilton WL, Leskovec J, Jurafsky D. Diachronic word embeddings reveal statistical laws of semantic change. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 2016. p. 1489–1501. https://doi.org/10.18653/v1/P16-1141

[10] Schlechtweg D, McGillivray B, Hengchen S, Dubossarsky H, Tahmasebi N. SemEval-2020 Task 1: Unsupervised lexical semantic change detection. In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. 2020. p. 1–23. https://doi.org/10.18653/v1/2020.semeval-1.1

[11] Hogan A, Blomqvist E, Cochez M, et al. Knowledge graphs. *ACM Computing Surveys*. 2021;54(4) 71. https://doi.org/10.1145/3447772

[12] Vrandečić D, Krötzsch M. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*. 2014;57(10):78–85. https://doi.org/10.1145/2629489

[13] Bian H. LLM-empowered knowledge graph construction: A survey. *arXiv* [Preprint]. 2025. arXiv:2510.20345. https://arxiv.org/abs/2510.20345

[14] Chen R, Jiang W, Qin C, et al. Are large language models effective knowledge graph constructors? *arXiv* [Preprint]. 2025. arXiv:2510.11297. https://arxiv.org/abs/2510.11297

[15] Lu Y, Liu Q, Dai D, et al. Unified structure generation for universal information extraction. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 2022. p. 5755–5772. https://doi.org/10.18653/v1/2022.acl-long.395

[16] Huguet Cabot PL, Navigli R. REBEL: Relation extraction by end-to-end language generation. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021. p. 2370–2381. https://doi.org/10.18653/v1/2021.findings-emnlp.204

[17] Wang X, Zhou W, Zu C, et al. InstructUIE: Multi-task instruction tuning for unified information extraction. *arXiv* [Preprint]. 2023. arXiv:2304.08085. https://arxiv.org/abs/2304.08085

[18] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Advances in Neural Information Processing Systems*. 2020;33. https://arxiv.org/abs/2005.11401

[19] Asai A, Wu Z, Wang Y, Sil A, Hajishirzi H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2024. https://arxiv.org/abs/2310.11511

[20] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT*. 2019. p. 4171–4186. https://doi.org/10.18653/v1/N19-1423

[21] Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: *Proceedings of EMNLP-IJCNLP*. 2019. p. 3982–3992. https://doi.org/10.18653/v1/D19-1410

[22] Lebo T, Sahoo S, McGuinness D, editors. PROV-O: The PROV ontology. *W3C Recommendation*. 2013. https://www.w3.org/TR/prov-o/

[23] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18

[24] Liu NF, Lin K, Hewitt J, et al. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*. 2024;12:157–173. https://doi.org/10.1162/tacl_a_00638

[25] Artstein R, Poesio M. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*. 2008;34(4):555–596. https://doi.org/10.1162/coli.07-034-R2

[26] Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. G-Eval: NLG evaluation using GPT-4 with better human alignment. In: *Proceedings of EMNLP*. 2023. p. 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153

[27] Lin CY. ROUGE: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. 2004. p. 74–81. https://aclanthology.org/W04-1013/