You Can Create Words Rather Than Just Detect Them: Towards A New Paradigm of Chinese Neologisms

Anonymous ACL submission

Abstract

Lexical evolution is universal across languages, yet the mechanisms driving neologism formation vary substantially from one language to another. In Chinese, this variation is particularly reflected in the productivity of compounding, where morphemes are combined through word-formation relations to create interpretable lexical forms. This morphological productivity implies that the Chinese lexicon can be viewed not only as a set of attested words, but also as a dynamically latent space with potential neologisms to be realized. However, this perspective still remains underexplored in computation, as current approaches are largely **reactive**, focusing on detecting attested words from corpora. Modeling such a latent space will shed light on earlier identification of plausible neologisms. Motivated by this, we therefore propose a complementary paradigm of Chinese neologism research through **proactive** creation. We introduce a three-stage pipeline: KG-Based Candidate Generation, Morphological Discarding, and LLM-Based Filtering. After the pipeline, an evaluation schema is designed to fully assess the qualification of generated candidates through human judgment, model perplexity, and corpus attestation. By establishing a computational framework to model the generative capacity of Chinese neologisms, our work reveals the existence of a vast, explorable latent lexical space, which will benefit exploration of both linguistic competence and computational capacity. Overall, this study substantiates a new paradigm of neologisms, which moves beyond reactively detecting what they *are* to proactively modeling what they *could be*.

1 Introduction

Human language is intrinsically dynamic: linguistic forms emerge, evolve, and fade over time (Weinreich et al., 1968). As societies and communicative practices change, so does the lexicon, where neologisms proliferate to accommodate novel concepts or functions (Algeo and Algeo, 1991; Alyeksyeyeva et al., 2020). In recent decades, cultural globalization and technological innovation, particularly the emergence of generative AI, have significantly accelerated this process (Čilić and Plauc, 2021; Yakura et al., 2025).

Although linguistic evolution is universal, the mechanisms that create new lexical forms are language-specific. This matters computationally because a model of word creation must first decide what its basic units are and what constraints govern their combination. Chinese is particularly suitable for such modeling: Compounding is a central and highly productive mechanism in Chinese, and modern Chinese strongly favors disyllabic lexical forms, many of which can be analyzed as two morphemes linked by word-formation types (Lu, 1964; Chao, 1965; Zhu, 1982; Duanmu, 1998; Packard, 2000). Together, these properties outlook a latent strucured space of plausible forms organized by morphemes and word-formation types.

This space also clarifies why corpus attestation alone cannot exhaust the phenomenon of lexical innovation. In terms of the competence–performance distinction, attested word forms reflect linguistic *performance*, whereas speakers' *competence* can license forms that are morphologically interpretable even before they become conventionalized (Chomsky, 1965; Aronoff, 1976). For example, a dictionary-absent disyllabic form may still be understandable when its morphemes and formation support a coherent interpretation, much as established compounds are interpreted compositionally (Wei et al., 2023; Hao et al., 2024). Under this view, the Chinese lexicon is not limited to observed entries; it also includes a latent space of plausible but unattested forms. It is large enough to contain many unattested possibilities, and also structured enough to be modeled computationally.

Existing computational work on Chinese neologisms has mainly treated new words as observed

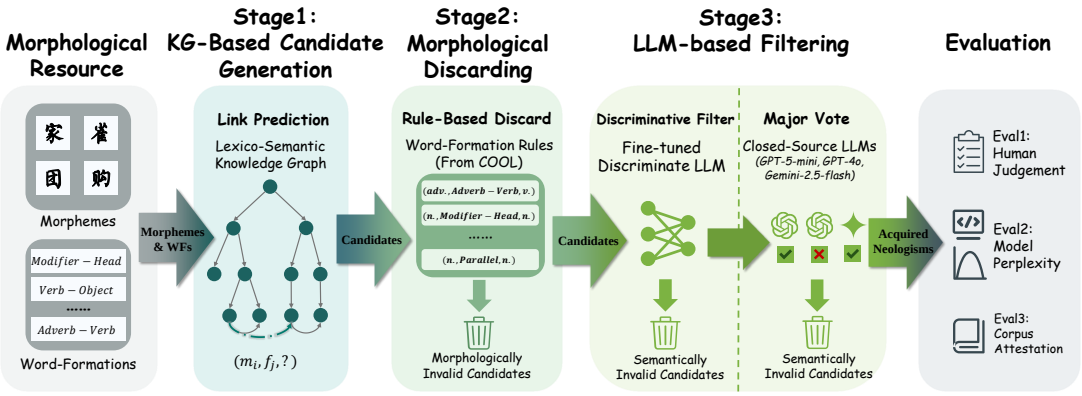

Figure 1: An illustration of our proposed pipeline with the evaluation scheme for Chinese neologism creation.

lexical events: a form first appears in text, and the system then identifies it using corpus statistics or distributional cues (Li et al., 2004; Peng et al., 2004; Pang et al., 2025). This detection-oriented framing is important for lexicon maintenance, downstream NLP systems, and the empirical study of lexical change. However, it leaves an earlier generative question largely unaddressed: before a form becomes reliably attested, can the structured morphology of Chinese be used to construct plausible lexical candidates for emerging meanings? It can therefore be formulated as **proactive neologism creation**: generating disyllabic candidates from a structured space of morphemes and word-formation relations before reliable corpus attestation is available.

To this end, we propose a three-stage pipeline that proactively creates Chinese neologisms: **Knowledge Graph Based Candidate Generation**, recalling plausible candidates via Link Prediction; **Morphological Discarding**, enforcing morphological constraints to eliminate invalid structures; **LLM-Based Filtering**, employing both fine-tuned and closed-source models to filter semantically invalid candidates.

To rigorously assess the output of our pipeline, we conduct a three-stage evaluation aligned with linguistic criteria: **Human Judgment**, serving as a holistic gold standard to assess morphological, semantic, and pragmatic validity; **Model Perplexity**, quantifying semantic coherence by measuring the contextual naturalness of candidates in language models; **Corpus Attestation**, verifying pragmatic plausibility by identifying emerging usage in large-scale corpora.

For illustration, in Figure 1, we provide a diagram of both the proposed pipeline and the evaluation scheme for Chinese neologism creation.

Experimental results substantiate that the neologisms created by our pipeline significantly outperform negative baselines morphologically, semantically and pragmatically. They exhibit validity comparable to existing words and maintain a decisive advantage in valid usage despite their lower frequency. These findings show that proactive generation can complement detection by modeling plausible lexical forms beyond the currently attested vocabulary. Consequently, this work provides a computational framework for exploring what Chinese word-formation rules allow, not only what corpora have already recorded.

In this paper, our main contributions are threefold[1]:

(1) We propose a new paradigm for **proactively creating Chinese neologisms**, realized by a three-stage pipeline that integrates Knowledge Graph based candidate generation, morphological discarding, and LLM-based filtering;

(2) We design an **evaluation scheme** combining human judgment, model perplexity and corpus attestation;

(3) Following the pipeline and the evaluation, a **high-quality dataset** of the generated neologisms is also released to facilitate research in lexical evolution.

2 Related Work

2.1 Definition and Acquisition of Neologisms

Defining a neologism is not straightforward because a word's point of origin is often difficult to trace (Algeo and Algeo, 1991). We therefore adopt the operational definition commonly used in computational work: a neologism is a word form absent from a predefined reference vocabulary (Algeo and

[1]The resources and code will be released to the public once the paper is accepted.

Algeo, 1991; Herbelot and Baroni, 2017; Zheng et al., 2024). Under this view, the reference vocabulary functions as an observed lexicon, and absent forms mark potential lexical expansion driven by the language's generative capacity (Lehrer, 2003; Kerremans, 2016).

Historically, neologism acquisition has been framed mainly as detection. In English, where new forms often appear as distinct orthographic strings, methods track Out-of-Vocabulary (OOV) tokens or frequency bursts (Renouf, 1993; Cook, 2011). In Chinese, by contrast, many neologisms are combinations of familiar characters without explicit word boundaries, making identification closely tied to segmentation (Chen and Ma, 2002); methods therefore rely on signals such as Pointwise Mutual Information (PMI) (Sproat and Shih, 1990) and Branching Entropy (Jin and Tanaka-Ishii, 2006). Recent work extends this detection-oriented view with contextual embeddings, noisy-text normalization, and downstream evaluations of unseen lexical forms (Herbelot and Baroni, 2017; Torres-Rivera and Torres-Moreno, 2020; Wei et al., 2020; Zalmout et al., 2019; Zheng et al., 2024). Taken together, this literature explains how new forms can be identified or processed after they have appeared in text, but says less about how the space of plausible unattested forms can be explored before reliable corpus evidence is available.

A small complementary line has begun to address this problem, but it remains narrow in scope. In English, existing work focuses primarily on portmanteaus (blending two words, e.g., *brunch*) (Gangal et al., 2017). In Chinese, the "parallel condition" serves as a guiding theoretical principle, positing that near-synonymous morphemes share similar combinatory potentials (Chen, 2009). Guided by this principle, early attempts used heuristic substitution on analogical templates to generate compounds (Kang et al., 2020; Wang, 2023). However, these approaches were often limited to surface-level analogies, failing to ensure either semantic coherence or pragmatic plausibility.

2.2 Chinese Morphological Resources and Compound Productivity

Theoretically, Chinese word formation is strongly shaped by compounding, where morphemes combine through word-formation types to construct lexical meanings (Lu, 1964; Chao, 1965; Packard, 2000). Prior corpus-based studies of Chinese morphology and neologisms further show that Mandarin compounds remain productive and exhibit multiple headedness patterns (Sproat and Shih, 1996; Shen and Baayen, 2022; Ceccagno and Basciano, 2007; Ceccagno, 2016). This literature motivates our scope: rather than modeling all neologism mechanisms, such as borrowing, abbreviation, semantic drift, or event-specific coinages, we focus on disyllabic compounds built from morphemes and word-formation types.

Grounded in these theories, computational resources typically provide morpheme-level definitions (Yuan and Huang, 1998; Kang et al., 2004) or word-level formation features (Yuan and Huang, 1998; Wang et al., 2023). The COOL Dataset (Liu et al., 2018) uniquely connects these granularities: it organizes morphemes under sense categories and annotates words with both their constituent morphemes and word-formation types. This unified organization has supported downstream applications such as definition generation (Zheng et al., 2021a) and word sense disambiguation (Zheng et al., 2021b; Wang et al., 2024). We therefore use COOL as the foundational resource for constructing our morpheme and formation-relation space.

3 Method

In this section, we formally define the task of Chinese disyllabic neologism creation. After that, as detailed in Figure 1, a three-stage pipeline is designed for the task: **KG-Based Candidate Generation, Morphological Discarding, and LLM-based Filtering**.

3.1 Task Definition

We formulate the creation of Chinese disyllabic neologisms as an exploration of the combinatorial space spanned by word-formations and morphemes. Given a morpheme vocabulary $\mathcal{M} = \{m_1, m_2, \ldots, m_n\}$, we define the word-formation inventory as $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$, where each f_j is one of the word-formation types, like those annotated in the COOL dataset (Liu et al., 2018). Each operator specifies the structural type between the first and second morpheme: for example, *Modifier-Head* combines a modifying morpheme with a nominal head, *Verb-Object* combines a verbal morpheme with its object, and *Parallel* coordinates two semantically similar or complementary morphemes. The space of all theoretically possible disyllabic word constructions is then defined as the Cartesian product $\mathcal{S} = \mathcal{M} \times \mathcal{F} \times \mathcal{M}$. An

element within this space is denoted as a triplet $w_{ijk} = (m_i, f_j, m_k)$. We posit that such a triplet represents a potential lexical item, where the same ordered morpheme pair can receive different interpretations under different word-formation types.[2]

Under this framework, the objective is to learn a validity function $V : \mathcal{S} \rightarrow \{0, 1\}$, which identifies a subset of valid neologisms $\mathcal{N} \subset \mathcal{S}$. A triplet w_{ijk} is deemed to belong to $\mathcal{N}$ if and only if it is novel (absent from the standard lexicon)[3] and satisfies three linguistic criteria: **Morphological Soundness:** The combination strictly adheres to the morphological constraints of modern Chinese word-formation rules; **Semantic Coherence:** The synthesis of morphemes m_i and m_k yields a meaningful and intelligible concept derived from its components; **Pragmatic Plausibility:** The word functions naturally within linguistic contexts, exhibiting traces of use to established lexicon.

In this study, we ground this definition using the COOL dataset (Liu et al., 2018), which provides approximately 20,000 morphemes and the word-formation types $\mathcal{F}$ (see Appendix A for the full taxonomy, examples, and distribution). To target neologisms with high communicative potential, the search space $\mathcal{M}' \subset \mathcal{M}$ is restricted to the top 3,000 most commonly used morphemes (sorted with the appearance in the MiCLS dataset (Wang et al., 2024)). This constrained vocabulary still yields $|\mathcal{M}'| \times |\mathcal{F}| \times |\mathcal{M}'| = 3{,}000 \times 16 \times 3{,}000 = 1.44 \times 10^8$ possible triplets, which motivates the high-recall KG retrieval stage described below. The standard lexicon is compiled from four authoritative sources in Chinese: *Tongyici Cilin* (Mei et al., 1983), *Grammatical Knowledge-base of Contemporary Chinese* (Yu et al., 1996), *HowNet* (Dong and Dong, 2003), and *The Contemporary Chinese Dictionary* (Dictionary Editing Office, 2016).

3.2 Stage 1: Knowledge Graph Based Candidate Generation

This stage builds a high-recall pool of candidates for later filtering. We describe morphemes and word-formations in a Knowledge Graph, so plausible unseen combinations can be retrieved through Link Prediction. After that, each candidate is enriched to facilitate downstream filtering and evaluation.

Knowledge Graph Construction. The Knowledge Graph is defined as $\mathcal{G} = (\mathcal{E}, \mathcal{R})$, where $\mathcal{E}$ denotes entities and $\mathcal{R}$ denotes relations (See Figure 2(a)). The entities, drawn from the COOL Dataset (Liu et al., 2018), comprise two types: **Morphemes** and **Morpheme Sense Categories**, where the latter are semantic classes that group morphemes by sense. The set of relations $\mathcal{R}$ captures three distinct linguistic properties: **Word-Formation:** Edges connecting two morphemes to represent an existing word, comprising the 16 relation types in $\mathcal{F}$, each corresponding to a specific word-formation operator; **Morpheme Hyponymy:** Edges linking morpheme sense categories to represent the taxonomic hierarchy; **Morpheme Instantiation:** Edges linking a category to a specific morpheme instance.

Link Prediction. We formulate candidate generation as Link Prediction: given a head morpheme m_i and word-formation f_j, retrieve plausible tail morphemes m_k to complete $(m_i, f_j, ?)$ (See Figure 2(b)). We instantiate this stage with SimKGC (Wang et al., 2022), using `bert-base-chinese` (Devlin et al., 2019) to encode morpheme definitions. SimKGC represents the head–relation pair (m_i, f_j) as a query vector $\mathbf{q}_{ij}$ and each tail morpheme m_k as an entity vector $\mathbf{e}_k$, yielding the relation-conditioned score:

$$\phi(m_i, f_j, m_k) = \cos(\mathbf{q}_{ij}, \mathbf{e}_k). \tag{1}$$

Because $\mathbf{q}_{ij}$ includes f_j, the same head morpheme can produce different ranked tail lists under different word-formations; Appendix D verifies this effect through an ablation study. We retain all candidates whose scores exceed a lenient validation-tuned threshold[4], reducing the raw 1.44×10^8 triplet space to 802,108 KG-retrieved candidates before more expensive filtering.

Contextual Enrichment. To facilitate downstream filtering and evaluation, each candidate is enriched with a generated definition and usage contexts. First, an advanced LLM, Claude-3.7-Sonnet (Anthropic, 2025), produces a definition

[2] While natural language allows for polysemy where one form maps to multiple meanings, for computational tractability, a **uniqueness assumption** is adopted here: each triplet w_{ijk} corresponds to at most one single dominant lexical entry. This allows us to model the generation task as a discrete validity classification problem.

[3] This criterion ensures the focus remains on proactive creation — discovering novel morpheme pairings rather than analyzing semantic drift in existing words.

[4] In our experiments, we set $\tau = 0.3$ based on analysis of the validation set to ensure novel yet plausible combinations are not prematurely filtered.

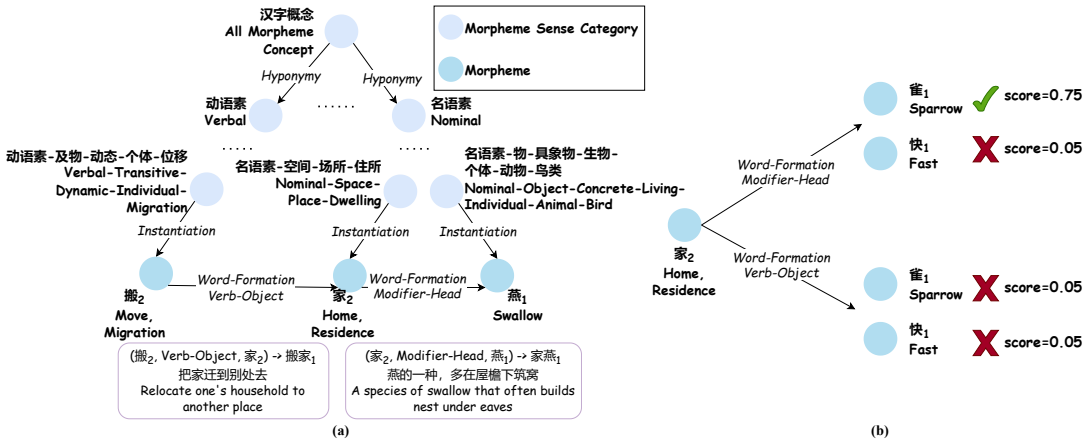

Figure 2: Illustration of the Candidate Generation on the Knowledge Graph: **(a) Knowledge Graph Construction:** The graph encodes the structural information inside Chinese morphology; **(b) Link Prediction via SimKGC:** Given a head morpheme (e.g., 家 "Home") and a target word-formation (e.g., *Modifier-Head*), the model predicts the most plausible tail morphemes. The checkmark (√) indicates a valid high-scoring neologism candidate ("家雀"), while crosses (×) denote invalid combinations.

from word-formation and morpheme meanings.[5] Subsequently, an ensemble of five LLMs are used to generates five sample sentences.[6] Crucially, during this process, models are prompted with the definition while masking the word form, which prevents orthographic-based hallucinations and ensures the generated contexts are grounded in the intended compositional meaning.

3.3 Stage 2: Morphological Discarding

Following the generation phase, we enforce morphological soundness with a rule-based filter. This approach is grounded in the principle that Chinese neologisms usually form by analogy to established words (Dong, 2005). The filter therefore discards any candidate whose pattern is unattested, rejecting morphologically invalid combinations.

Morpheme POS tags (grammatical categories for morphemes analogous to word-level POS), extracted from the COOL Dataset (Liu et al., 2018), are leveraged for the rule-based filter. From these tags, a whitelist of all morphological patterns observed in attested words is compiled, formalized as $(HeadMorphemePOS, Word-Formation, TailMorphemePOS)$ triplets. Any candidate created by Stage 1, whose pattern does not appear in this whitelist, is subsequently discarded. This whitelist retains 73,381 triplets; after surface-form de-duplication, 46,998 candidates are passed to Stage 3.

3.4 Stage 3: LLM-based Filtering

The final stage focuses on ensuring high semantic quality. We implement a two-tier filtering process: coarse-grained screening with a fine-tuned discriminative LLM, followed by fine-grained verification with stronger LLM-based judges.

Discriminative LLM Filtering. To efficiently process the remaining candidates, we train a specialized binary discriminative model, with the Qwen3-Embedding-4B model (Zhang et al., 2025) as the backbone and an MLP as the classifier head. The embedding backbones used in Stage 1 and Stage 3 are summarized in Appendix B. The training data, derived from the COOL Dataset (Liu et al., 2018), is enriched via a sophisticated **data augmentation strategy** designed to encourage abstract semantic reasoning over surface-level patterns. We construct positive samples not only from existing words but also via *Category Substitution* (replacing morphemes with their hypernyms) and *Sibling Substitution* (simulating valid analogies); Conversely, negative samples are generated with varying difficulty levels (including random, morpheme POS constrained, and high-frequency substitution) to refine the decision boundary (see Appendix E for detailed implementation). The model is optimized with a joint Cross-Entropy and Su-

[5]For computational and analytical convenience, we assume here that word meaning is composed directly from word-formation and morpheme meaning, without considering additional factors such as semantic drift. This conforms to the formation pattern of most neologisms in Chinese (Dong, 2005).

[6]They are Gemini-2.0-flash (Google, 2025), Gemini-2.5-flash (Comanici et al., 2025), GPT-4o-mini (Hurst et al., 2024), GPT-4.1-mini (OpenAI, 2025b) and GPT-5-mini (OpenAI, 2025a), chosen for their superior ability and efficiency.

pervised Contrastive Loss ($\mathcal{L}\text{total} = \mathcal{L}\text{CE} + 0.5 \cdot \mathcal{L}_{\text{SupCon}}$) to shape a semantically discriminative latent space. Only candidates classified as positive proceed to the final verification.

Verification via Closed-Source LLM. To eliminate subtle semantic inconsistencies that the discriminative model might miss, we employ a verification mechanism using three distinct LLMs as expert annotators: GPT-4o (Hurst et al., 2024), GPT-5-mini (OpenAI, 2025a), and Gemini-2.5-Flash (Comanici et al., 2025). Each model makes a binary judgment on each candidate's plausibility, considering its word-formation, morphemes, and the definition and sentences generated in Stage 1 (see Appendix F for prompt). A candidate is accepted upon receiving at least two positive votes (Majority Vote $\geq$ 2). This ensemble approach leverages diverse LLMs to impose a rigorous standard for semantic soundness.

Overall, the pipeline reduces the enumerated space to 18,662 final positive candidates; Appendix C provides the full stage-wise funnel.

4 Evaluation

To validate our pipeline, we setup our experiment, and then evaluate candidates via **Human Judgment**, **Model Perplexity**, and **Corpus Attestation**. These evaluations test whether candidates passing all filtering stages are substantially more plausible than rejected candidates and approach established words in validity; Appendix G provides a detailed statistical breakdown of the evaluation sets.

4.1 Experimental Setup

Three distinct sets are constructed for evaluation:

Positive Samples (POS): The set of **18,662** candidates that successfully passed all three stages of the pipeline;

Negative Samples (NEG): The set of **590,011** candidates that passed the initial Knowledge Graph screening but were rejected by either the rule-based or the LLM-based filter[7];

Existing Words (EXIST): A baseline set of **42,844** established, common disyllabic words sampled from COOL (Liu et al., 2018), with word-formation and morpheme annotations. To ensure comparability, we generate sample sentences for them using the same LLM-based method described in Section 3.2.

4.2 Human Judgment

To assess *real-world acceptability*, we randomly sample 200 POS and 200 NEG words for **Human Judgement Evaluation**. Three computational linguistics experts judge each item using its word form, word-formation, morpheme senses, LLM-generated definition, and LLM-generated sentences, assigning a binary acceptance label under our three criteria: *morphological soundness*, *semantic coherence*, and *pragmatic plausibility*. Appendix H provides the annotation instructions.

We hypothesize that the acceptance rate for POS samples will be significantly higher than for NEG samples: ACC(POS) $>$ ACC(NEG), demonstrating the efficacy of the pipeline in identifying high-quality candidates.

4.3 Model Perplexity

To quantify *Semantic Coherence*, **Model Perplexity** measures each target word's contextual plausibility in its five generated context sentences (Section 3.2). We evaluate PPL with both Bidirectional Masked Language Models and Causal Autoregressive Models.[8] Given the distinct architectures, specific calculation protocols are adopted (details in Appendix I):

For **Bidirectional Models**, we compute PPL by iteratively masking each character of $w = (c_1, c_2)$ and exponentiating the average negative log-likelihood of reconstructing it from the bidirectional context; For **Causal Models**, since standard next-token prediction is inadequate for infill evaluation, we use a cloze task: masking the target word, appending a query (e.g., "*The word here refers to:*"), and exponentiating the average autoregressive negative log-likelihood of c_1 and c_2. Scores are averaged across three query templates to mitigate prompt sensitivity.

We use a **Mann-Whitney U test** to compare POS against NEG and EXIST distributions and report the p-values. Since semantically valid neologisms should yield lower PPL than invalid

[7]The NEG set includes candidates with plausible morpheme pair but invalid word-formation. Since Model Perplexity and Corpus Attestation are blind to this structural difference, for these two evaluations, we exclude any NEG candidate whose word form also appears in the POS set to avoid confounding the results.

[8]BERT-Base-Chinese (Devlin et al., 2019) and ModernBERT-Large (Warner et al., 2025) are used as masked language models, while Llama-3.2-3B (Meta AI, 2024) and Qwen3-1.7B (Yang et al., 2025) are used as autoregressive models, chosen for a robust evaluation across diverse architectures and scales.

Set	Avg. Acc. Rate	Annotator Rates	Kappa
POS	69.83	71.0, 66.5, 72.0	0.6202
NEG	26.67	26.5, 23.5, 30.0	0.6250

Table 1: Human judgment results: average acceptance rate (%), annotator-specific rates (%), and Fleiss' Kappa.

combinations, we hypothesize: PPL(NEG) $>$ PPL(POS) $>$ PPL(EXIST).

4.4 Corpus Attestation

To verify *Pragmatic Plausibility*, **Corpus-Based Attestation** measures the extent to which candidates are already emerging by searching all POS, NEG, and EXIST words in a large-scale Chinese corpus[9]. Because raw string matching yields false positives (e.g., from homographs) and traditional WSD models fail on Out-of-Distribution neologisms, we use LLM-based semantic verification. For each set, 10,000 retrieved context instances are sampled, and *Gemini-2.5-Pro* (Comanici et al., 2025) verifies whether the contextual usage aligns with the definition. The resulting validity ratio is extrapolated to all retrieval results to estimate total valid attestations.

We anticipate that the number of valid attestations will follow the order: Count(EXIST) $\gg$ Count(POS) $>$ Count(NEG), where the count for existing words serves as a strong upper bound.

5 Results and Analysis

5.1 Results of Human Judgment

The results of Human Judgment (As illustrated in Table 1) demonstrate the pipeline's effectiveness. The acceptance rate for the POS set (69.83%) is significantly higher than for the NEG set (26.67%), confirming the capability of identifying high-potential candidates. The inter-annotator agreement (Fleiss' Kappa $\approx$ 0.62) indicates substantial agreement (Landis and Koch, 1977), a result that reflects the inherent subjectivity of evaluating neologisms. Additional bootstrap and attestation-stratified analyses are provided in Appendices K and L.

5.2 Results of Model Perplexity

The results of Model Perplexity (As illustrated in Table 2) validate the hypothesis: PPL(NEG) $>$ PPL(POS) $>$ PPL(EXIST), with all adjacent groups showing statistically significant differences (Mann-Whitney U test, $p < 0.01$). Although existing words achieve the lowest PPL, the gap with the POS set is remarkably narrow on stronger models, confirming the candidates' high contextual plausibility (see Appendix M for a detailed breakdown).

The density distributions in Figure 3 reinforce these findings. The distribution of POS set eliminates the high-perplexity tail of the NEG set and substantially overlaps with the distribution of EXIST set, suggesting many of the generated neologisms are as plausible as established words in semantic coherence.

[9] https://github.com/mobvoi/seq-monkey-data/blob/main/docs/pretrain_open_corpus.md

Model	POS		NEG		EXIST	
	Ave	Var	Ave	Var	Ave	Var
BERT-Base-Chinese	7.77	4.48	9.50	4.13	4.28	8.62
ModernBERT-large	2.65	2.98	3.41	4.37	2.47	3.04
Llama-3.2-3B	8.79	5.06	9.48	4.16	7.82	10.12
Qwen3-1.7B	9.61	3.30	10.88	3.15	7.82	4.50
Average	7.20	1.85	8.32	1.77	5.62	2.61

Table 2: Model perplexity results for POS, NEG, and EXIST sets (PPL; lower is better).

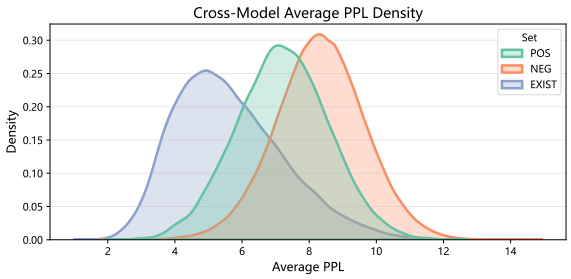

Figure 3: Density distribution of average PPL scores across models.

5.3 Results of Corpus Attestation

The results of Corpus Attestation (As illustrated in Table 3) confirm the hypothesis: Count(EXIST) $\gg$ Count(POS) $>$ Count(NEG). The candidates in POS set achieve a validity ratio of 18.3% (over four times higher than the NEG set's 4.0%), translating to an estimated 536k valid occurrences versus just 87k. While lagging behind the frequency of established words ($\sim$21.8M), this substantial presence confirms that the pipeline effectively identifies high-potential neologisms already emerging in the long tail of language. Appendix J further separates POS candidates by corpus attestation to distinguish attested emerging forms from unattested latent candidates.

Category	Total Lines	Sampled	Valid Count	Validity Ratio (95% CI)	Est. Valid Total
EXIST	50,346,797	10,000	4,329	0.433 (±0.010)	21,795,128
POS	2,930,062	10,000	1,831	0.183 (±0.008)	536,494
NEG	2,202,294	10,000	395	0.040 (±0.004)	86,991

Table 3: LLM-based semantic verification for corpus attestation, with 95% confidence intervals.

Neologism	POS	WF	Generated Definition
团购 (Group Buying)	V	Adverb-Verb	Organize a group to buy goods.
司命 (Arbiter of Fate)	N	Verb-Object	The power that controls destiny and life.
尘外 (Beyond Mundane)	N	Entity-Position	A realm beyond the mortal, "dusty" world.

Table 4: Case-study examples with POS tags, word-formation types, and generated definitions.

5.4 Qualitative Analysis

The cases in Table 4 illustrate two types of candidates identified by the pipeline. **团购** (Group Buying) and **司命** (Arbiter of Fate) are already attested in the corpus, suggesting emerging real-world usage. In contrast, **尘外** (Beyond Mundane) remains unattested but was judged valid by human annotators, indicating that the pipeline can also uncover plausible latent forms. Appendix N analyzes representative disagreement cases between automatic labels and human judgments. A more detailed case study is provided in Appendix O.

5.5 Overall Discussion

Overall, our evaluation schema provides three main insights into the automated generation of Chinese neologisms:

The triangulation of evidence across Human Judgement, Model Perplexity and Corpus Attestation provides **robust validation** for the pipeline. The consistent ranking, EXIST $>$ POS $>$ NEG, across all metrics confirms that our method captures implicit morphological, semantic and pragmatic constraints in Chinese;

The evaluation supports the view that Chinese word formation functions as a **lexical continuum**. Chinese compounds do not fall into a simple valid/invalid dichotomy but occupy a gradient of lexicalization (Hopper and Traugott, 2003; Dong, 2005). Our method's probabilistic scoring mirrors this property by assigning quantifiable plausibility scores, thereby modeling the lexicon as a continuous space and capturing the generative competence of Chinese words;

The pipeline also reveals **the lexicon's latent space**. Beyond detecting neologisms already attested in corpora, it identifies high-quality latent words: linguistically plausible yet statistically unseen candidates. This substantiates the view of the lexicon not as a static repository, but as a dynamic ecosystem where novel concepts can be systematically lexicalized.

6 Conclusion

In this paper, we explore Chinese neologism from a proactive perspective. Rather than waiting for new words to become attested and then detecting them, we ask whether Chinese word-formation knowledge can be used to construct plausible disyllabic forms in advance. To answer this question, we propose a three-stage pipeline that integrates KG-based candidate generation, morphological discarding, and LLM-based filtering. We evaluate the generated forms through human judgment, model perplexity, and corpus attestation, which together assess their morphological validity, semantic coherence, and pragmatic plausibility. The results suggest that the pipeline does more than produce isolated candidates: it provides a computational framework for exploring a latent lexical space of plausible forms licensed by Chinese word-formation rules, beyond what corpora have already recorded.

In the future, further research is to expand in the following directions: The paradigm of proactive creation offers a valuable reference for exploring the generative potential in other languages with their corresponding resources; Additionally, temporal dynamics could be incorporated by analyzing diachronic corpora to model how the plausibility of a neologism evolves alongside shifting cultural and technological contexts; Finally, research in LLMs could leverage these neologisms to enhance the models' core linguistic capabilities, like improving the understanding of lexical novelty and enriching the capacity for Creative Generation.

Limitations

Our work is subject to several limitations:

Scope of Neologism Type: Our pipeline focuses exclusively on orthographically novel disyllabic

neologisms, overlooking other forms of lexical innovation like polysemy (new meanings for existing words), acronyms, or loanwords;

Depth of Semantic Modeling: The semantic modeling is constrained to compositional meanings directly derivable from morphemes. It does not capture non-compositional semantics such as metaphor, metonymy, or irony;

Inherent Biases in Evaluation: Each evaluation metric has inherent limitations. Human judgments can be subjective, model perplexity is biased by training data, and corpus attestation is a lagging indicator that cannot assess truly novel "zero-shot" neologisms. Indeed, establishing a truly effective and comprehensive evaluation methodology for this task remains an open research topic.

References

John Algeo and Adele S Algeo. 1991. *Fifty years among the new words: A dictionary of neologisms 1941-1991*. Cambridge University Press.

Iryna O Alyeksyeyeva, Tetyana A Chaiuk, and Elizaveta A Galitska. 2020. Coronaspeak as key to coronaculture: Studying new cultural practices through neologisms. *International Journal of English Linguistics*, 10(6):202–212.

Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet. Accessed: 2025-11-13.

Mark Aronoff. 1976. Word formation in generative grammar. *Linguistic Inquiry Monographs Cambridge, Mass*, (1):1–134.

Antonella Ceccagno. 2016. Chinese neologisms: Word-formation strategies in chinese. In *The Routledge encyclopedia of the Chinese language*, pages 265–279. Routledge.

Antonella Ceccagno and Bianca Basciano. 2007. Compound headedness in chinese: An analysis of neologisms. *Morphology*, 17(2):207–231.

Yuen Ren Chao. 1965. A grammar of spoken chinese.

Baoya Chen. 2009. *Contemperary Linguistics*. Higher Education Press, Beijing, China.

Keh-Jiann Chen and Wei-Yun Ma. 2002. Unknown word extraction for chinese documents. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. 1. MIT press.

Ivona Šetka Čilić and Jelena Ilić Plauc. 2021. Today' s usage of neologisms in social media communication. *Društvene i humanističke studije*, 6(1 (14)):115–140.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

C Paul Cook. 2011. *Exploiting linguistic knowledge to infer properties of neologisms*. Ph.D. thesis.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Chinese Academy of Social Sciences Dictionary Editing Office, Institute of Linguistics. 2016. *Contemporary Chinese Dictionary (Seventh Edition)*. The Commercial Press, Beijing, China.

Xiufang Dong. 2005. *Lexicon and Grammar of Chinese*. Peking University Press, Beijing, China.

Zhendong Dong and Qiang Dong. 2003. Hownet-a hybrid language and knowledge resource. In *International conference on natural language processing and knowledge engineering, 2003. Proceedings. 2003*, pages 820–824. IEEE.

San Duanmu. 1998. Wordhood in chinese. *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese*, pages 135–196.

Aiping Fu. 2003. Word formation and the recognition of compounds in chinese language understanding. *Applied Linguistics*, 4:25–33.

Varun Gangal, Harsh Jhamtani, Graham Neubig, Eduard Hovy, and Eric Nyberg. 2017. Charmanteau: Character embedding models for portmanteau creation. *arXiv preprint arXiv:1707.01176*.

Google. 2025. Gemini 2.0: Flash, Flash-Lite and Pro. https://developers.googleblog.com/en/gemini-2-family-expands/. Accessed: 2025-11-13.

Shaoyu Guo. 1979. *New Exploration into Chinese Grammatical Rhetoric*. The Commercial Press, Beijing, China.

Yuxin Hao, Chenxi Wu, and Xun Duan. 2024. Processing mandarin chinese compound words by native speakers and second language learners: word frequency, semantic transparency, and word structure. *Sage Open*, 14(2):21582440241256249.

Aurélie Herbelot and Marco Baroni. 2017. High-risk learning: acquiring new word vectors from tiny data. *arXiv preprint arXiv:1707.06556*.

Paul J Hopper and Elizabeth Closs Traugott. 2003. *Grammaticalization*. Cambridge university press.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zhihui Jin and Kumiko Tanaka-Ishii. 2006. Unsupervised segmentation of chinese text by use of branching entropy. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 428–435.

Shiyong Kang, Yi Li, and Daogong Sun. 2004. Construction of chinese system corpus and dictionary compilation. In *Proceedings of the 2004 Workshop on Lexicography and Digitalization*.

Sichen Kang, Mengxia Yu, and Liu Yang. 2020. Knowledge representation and prediction of chinese unknown words via parallel conditions. *Journal of Chinese Information Processing*, 34(8):23–31.

Daphné Kerremans. 2016. 6. a web of new words. a corpus-based study of the conventionalization process of english neologisms. *English and American Studies in German*, 5(1):8–10.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Adrienne Lehrer. 2003. Understanding trendy neologisms. *Italian Journal of Linguistics*, 15:369–382.

Hongqiao Li, Chang-Ning Huang, Jianfeng Gao, and Xiaozhong Fan. 2004. The use of svm for chinese new word identification. In *International Conference on Natural Language Processing*, pages 723–732. Springer.

Shuxin Liu. 1990. *Chinese Descriptive Lexical Studies*. Commercial Press, Beijing, China.

Yang Liu, Zi Lin, and Sichen Kang. 2018. Towards a description of chinese morphemic concepts and semantic word-formation. *Journal of Chinese Information Processing*, 32(2):11–20.

Zhiwei Lu. 1964. *Word-Formation in Chinese*. The Scientific Press, Beijing, China.

Jiajv Mei, Yiming Zhu, Yunqi Gao, and Hongxiang Yin. 1983. *Chinese Cilin*. Shanghai Lexicographical Publishing House.

Meta AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Accessed: 2025-12-12.

OpenAI. 2025a. GPT-5 System Card. Technical report, OpenAI. Accessed: 2025-08-10.

OpenAI. 2025b. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. Accessed: 2025-12-3.

Jerome L Packard. 2000. *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge University Press.

Xian Pang, Chen Bo, and Zhao Xiaobing. 2025. Extraction and analysis of chinese-tibetan new words from news texts. *Acta Scientiarum Naturalium Universitatis Pekinensis*, pages 45–52.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 562–568.

Antoinette Renouf. 1993. A word in time: first findings from the investigation of dynamic text. *English Language Corpora: Design, Analysis and Exploitation, Rodopi, Amsterdam*, pages 279–288.

Tian Shen and R. Harald Baayen. 2022. Productivity and semantic transparency: An exploration of word formation in mandarin chinese. *The Mental Lexicon*, 17(3):458–479.

Richard Sproat and Chilin Shih. 1990. A statistical method for finding word boundaries in chinese text. *Computer Processing of Chinese & Oriental Languages*, 4(4):336–351.

Richard Sproat and Chilin Shih. 1996. A corpus-based analysis of mandarin nominal root compound. *Journal of East Asian Linguistics*, 5(1):49–71.

Andrés Torres-Rivera and Juan-Manuel Torres-Moreno. 2020. Detecting new word meanings: A comparison of word embedding models in spanish. *arXiv preprint arXiv:2001.05285*.

Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022. Simkgc: Simple contrastive knowledge graph completion with pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4281–4294.

Mengxiang Wang. 2023. *Research on Automatic Generation of Chinese Words*. 1. Xueyuan press.

Yue Wang, Qiliang Liang, Yaqi Yin, Hansi Wang, and Yang Liu. 2024. Disambiguate words like composing them: A morphology-informed approach to enhance chinese word sense disambiguation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15354–15365.

Yue Wang, Yang Liu, Qiliang Liang, and Hansi Wang. 2023. Construction of chinese semantic word-formation and its computing applications. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Tingxin Wei, Wenlei Bai, and Weiguang Qu. 2020. Sense prediction for chinese oov based on word embedding and semantic knowledge. *Data Analysis and Knowledge Discovery*, 4(6):109–117.

Yanjun Wei, Ying Niu, Marcus Taft, and Manuel Carreiras. 2023. Morphological decomposition in chinese compound word recognition: Electrophysiological evidence. *Brain and Language*, 241:105267.

Uriel Weinreich, William Labov, and Marvin Herzog. 1968. *Empirical foundations for a theory of language change*. University of Texas Press.

Tongqiang Xu. 1997. Core words and semantic word formation in chinese. *Linguistic Research*, 3:1–15.

Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, Ivan Soraperra, and Iyad Rahwan. 2025. Empirical evidence of large language model's influence on human spoken communication. *Preprint*, arXiv:2409.01754.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shiwen Yu, Xuefeng Zhu, Hui Wang, and Yunyun Zhang. 1996. Specification of grammatical knowledge-base of contemporary chinese. *Journal of Chinese Information Processing*.

Chunfa Yuan and Changning Huang. 1998. Research on chinese morphemes and word formation based on morpheme database. *Chinese Teaching in the World*.

Nasser Zalmout, Kapil Thadani, and Aasish Pappu. 2019. Unsupervised neologism normalization using embedding space mapping. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 425–430, Hong Kong, China. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Hua Zheng, Damai Dai, Lei Li, Tianyu Liu, Zhifang Sui, Baobao Chang, and Yang Liu. 2021a. Decompose, fuse and generate: A formation-informed method for chinese definition generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5524–5531.

Hua Zheng, Lei Li, Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, and Yang Liu. 2021b. Leveraging word-formation knowledge for chinese word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 918–923.

Jonathan Zheng, Alan Ritter, and Wei Xu. 2024. NEOBENCH: Evaluating robustness of large language models with neologisms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13885–13906, Bangkok, Thailand. Association for Computational Linguistics.

Dexi Zhu. 1982. *Grammar Lectures*. The Commercial Press, Beijing, China.

Yan Zhu. 2003. *Semantic Word-Formation Analysis of Modern Chinese*. Ph.D. thesis, East China Normal University.

A Details of Chinese Word-Formation

Word-formation refers to the linguistic rules and patterns by which morphemes are combined to form words. In Chinese, it involves the analysis of the internal structure of words, specifically, how morphemes combine through certain grammatical and semantic relationships to constitute a word. It operates as a structural system at the word level (Packard, 2000). Research shows that the internal structure of Chinese words involves not only morphemic composition but also semantic functions of constituents and meaning-generation mechanisms (Chao, 1965). Therefore, word-formation analysis is a crucial tool for understanding the Chinese lexical system.

Chinese word-formation can be broadly categorized into compounding, derivation, overlapping, etc., with compounding being the most predominant method (Chao, 1965).

Regarding the theoretical framework, there are different perspectives in linguistics, primarily distinguishing between **grammatical word-formation** and **semantic word-formation**. The grammatical word-formation classifies relationships using syntactic labels (e.g., *modifier-head*,

subject-predicate), arguing for a direct correspondence between sentence and word-formation types (Guo, 1979; Zhu, 1982). Conversely, the semantic word-formation emphasizes analyzing constituents through semantic labels (e.g., *agent*, *patient*), positing that characters combine into groups based on these semantic relationships (Liu, 1990; Xu, 1997; Zhu, 2003).

While semantic word-formation has natural advantages in representing lexical semantics, its structural basis is too complex to establish a unified tag set, making it less conducive to computational processing (Fu, 2003; Wang et al., 2023). In contrast, grammatical word-formation offers a simple structural system with unified standards, and the natural similarity between morphological and syntactic structures makes it more suitable for computational handling. Adopting the grammatical word-formation perspective, the COOL Dataset (Liu et al., 2018) establishes a taxonomy of 16 distinct word-formation types. In the main formulation, this taxonomy is denoted as $\mathcal{F} = \{f_1, \ldots, f_{16}\}$. Each $f_j \in \mathcal{F}$ is a grammatical operator over an ordered morpheme pair (m_i, m_k), specifying how the two morphemes combine into a candidate disyllabic word. For instance, the *Modifier-Head* operator treats the first morpheme as a modifier and the second as the head; the *Verb-Object* operator treats the first morpheme as an action and the second as its object; and the *Parallel* operator combines two morphemes with similar, coordinated, or complementary meanings. All 16 word-formation operators are illustrated in Table 5.

B Implementation Details of Embedding Models

The pipeline uses different embedding models for different roles. Stage 1 uses `bert-base-chinese` as the character-embedding backbone in SimKGC link prediction, where the model ranks candidate tail morphemes conditioned on the head morpheme and word-formation operator. Stage 3 uses Qwen3-Embedding-4B as the semantic embedding backbone for the discriminative filtering model, followed by an MLP classifier trained to distinguish plausible from implausible candidate words. Table 6 summarizes these implementation choices.

C Search Space and Candidate Funnel

This section details how the pipeline narrows the constrained search space introduced in Section 3. Using the top 3,000 morphemes and the 16 word-formation operators in $\mathcal{F}$ yields $3{,}000 \times 16 \times 3{,}000 = 1.44 \times 10^8$ possible morpheme–relation–morpheme triplets. The KG stage first retrieves a high-recall pool with threshold $\tau = 0.3$. Subsequent filters enforce lexical novelty, affix-position constraints, attested POS–formation compatibility, surface-form uniqueness, and semantic plausibility.

D Relation Ablation for KG-Based Candidate Generation

We evaluate whether the word-formation operator f_j contributes to the SimKGC retrieval step. The full model scores each candidate tail with the relation-conditioned function $\phi(m_i, f_j, m_k)$ in Equation 1, where the query is constructed from both the head morpheme and the target word-formation operator. The ablated model removes f_j from the query and ranks tails using only the head morpheme. Both variants are evaluated on held-out existing COOL triplets by ranking the gold tail morpheme among candidate tails.

E Training Details of Semantic Filtering Model

In this section, the comprehensive experimental setup for the Discriminative LLM Filtering procedure are provided, detailing the data construction strategy, hyperparameters, and prompt templates utilized in the experiments.

E.1 Data Construction with Semantic Augmentation

To train a robust discriminative model capable of generalizing to unseen neologisms, a train and validation set derived from COOL (Liu et al., 2018) is constructed, using a sophisticated augmentation strategy. A trivial classifier trained only on existing words might overfit to specific character collocations. To mitigate this, a multi-tiered positive and negative sampling scheme is designed. Table 9 provides concrete examples for each augmentation type based on the source word "**季春**" (*Last Spring Month*, where "季" means "The third month of a season", and "春" means "Spring").

Word-Formation Rule	Explanation	Use Case	%
定中 (Modifier-Head)	$morph_1$ modifies $morph_2$ (noun).	红花 (red-flower)	38.62
联合 (Parallel)	$morph_1$ and $morph_2$ are similar or complementary.	昏花 (dizzy-dim)	22.87
述宾 (Verb-Object)	$morph_1$ operates on $morph_2$.	花钱 (spend-money)	16.44
状中 (Adverb-Verb)	$morph_1$ modifies $morph_2$ (verb).	白花 (vainly-spend)	8.45
单纯 (Single Morpheme)	The word is a single morpheme.	花生 (peanut)	3.51
连谓 (Verb-Consequence)	$morph_2$ is the consequence of $morph_1$.	休息 (stop-rest)	3.43
后缀 (Suffixation)	$morph_2$ is the suffix of $morph_1$.	花头 (trick-∅)	2.70
述补 (Verb-Complement)	$morph_2$ is the action follows $morph_1$.	压低 (press-down)	1.28
主谓 (Subject-Predicate)	$morph_1$ is the subject of $morph_2$.	眼花 (eyesight-dim)	1.06
重叠 (Overlapping)	$morph_1$ and $morph_2$ are the same.	白白 (vainly-vainly)	0.59
方位 (Entity-Position)	$morph_1$ is an entity, $morph_2$ is a position.	期中 (semester-mid)	0.37
介宾 (Preposition-Object)	$morph_1$ is a preposition, $morph_2$ is an object.	凭空 (from-nowhere)	0.31
名量 (Noun-Quantifier)	$morph_2$ is the quantifier of $morph_1$.	花朵 (flower-bud)	0.13
数量 (Number-Quantifier)	$morph_1$ is a number, $morph_2$ is a quantifier.	一点 (one-dot)	0.10
前缀 (Prefixation)	$morph_1$ is the prefix of $morph_2$.	老师 (∅-teacher)	0.10
复量 (Quantifier-Quantifier)	Both $morph_1$ and $morph_2$ are quantifiers.	千米 (kilo-meter)	0.03

Table 5: The taxonomy of 16 Chinese word-formations from the COOL Dataset (Liu et al., 2018). Each word-formation is presented with its definition, a representative example, and its statistical prevalence in the lexicon.

Stage	Model	Role
Stage 1	bert-base-chinese	Character-embedding backbone for SimKGC link prediction.
Stage 3	Qwen3-Embedding-4B	Semantic embedding backbone for the discriminative MLP filter.

Table 6: Embedding models used in the two learned components of the pipeline. Stage 1 focuses on relation-conditioned KG retrieval, while Stage 3 focuses on semantic plausibility filtering.

Positive Samples (Valid Words). Expanded beyond existing words to encourage generalization, the positive set consists of:

Existing Words: Standard valid compounds attested in the lexicon (e.g., **季春**, "Last Spring Month");

Category Substitution: To force the model to learn abstract semantic compatibility rather than surface forms, we randomly replaced a morpheme with its corresponding *semantic category* defined in the KG. For instance, replacing "**春**" (*Spring*) with its semantic category "<*Season*>", yielding the semi-abstract form **季 <Season>**;

Sibling Substitution: We replaced a morpheme with a random sibling node from the same semantic category to simulate valid analogies. For example, substituting "**春**" (*Spring*) with "**夏**" (*Summer*) to form **季夏** ("Last Summer Month"). Since "**夏**" is a valid sibling of "**春**" under the "*Season*" category, the resulting word preserves the semantic logic of "The third month of [Season]".

The sampling ratio of these three positive types was set to **1:3:2**.

Negative Samples (Invalid Combinations). Negative samples are constructed with varying degrees of difficulty to refine the model's decision boundary:

Random Selection: Replacing a morpheme with another random-selected morpheme from the COOL Dataset. For example, replacing "**春**" (*Spring*) with the numeral "三" (*Three*) to form **季三** ("Last Three"), which is semantically distinct and nonsensical in this context.

POS-constrained Selection: Replacing a morpheme with one sharing the same morpheme POS. This forces the model to distinguish nuanced semantic incompatibility within the same semantic category. For example, replacing "**春**" (Noun) with "**水**" (*Water*, Noun) to form **季水** ("Last Water"). Although structurally valid (Noun-Noun), the temporal prefix "Last month of..." cannot logi-

Stage	Remaining	Raw %
Raw enumerated triplets	144,000,000	100.000
KG retrieval ($\tau = 0.3$)	802,108	0.557
Rule-filter input pool	165,102	0.115
After removing existing forms	99,340	0.069
After affix-position constraints	98,616	0.068
After POS–formation whitelist	73,381	0.051
After surface-form de-duplication	46,998	0.033
After LLM majority vote	18,662	0.013

Table 7: Candidate funnel from the constrained search space to the final positive set. "Raw %" is computed relative to the 1.44×10^8 enumerated morpheme–relation–morpheme triplets.

Retrieval Condition	Hit@1 (%)	Hit@3 (%)
Relation-conditioned SimKGC $\phi(m_i, f_j, m_k)$	73.0	86.6
Without relation condition $\phi(m_i, m_k)$	56.0	72.4

Table 8: Relation ablation for KG-based candidate generation. Removing the word-formation operator substantially reduces tail-retrieval accuracy, showing that f_j provides predictive structural information rather than serving only as a post-hoc label.

cally modify the substance "Water".

High-Frequency Selection: Replacing a morpheme with a high-frequency common morpheme. This prevents the model from biasing towards common characters regardless of meaning. For instance, replacing "**春**" with the high-frequency pronoun "**我**" (*Me*) to form **季我** ("Last Me"), breaking the semantic constraint.

The ratio of the three negative types was set to **2:1:1**. Crucially, we maintained a positive-to-negative ratio of **3:8**, a strictly imbalanced setting designed to impose a rigorous acceptance standard for neologisms.

E.2 Experimental Setup and Hyperparameters

Dataset Split. The augmented dataset is split into training and validation sets with a ratio of 9:1. The final training set consists of **1,352,832** samples, and the validation set contains **150,315** samples.

Computational Resources. The model was fine-tuned on **8 NVIDIA H20 GPUs**. The total duration for training and validation was approximately **85 hours**.

Hyperparameters. As described in Section 3.4, the model is optimized using a combined loss function $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \alpha \cdot \mathcal{L}_{\text{SupCon}}$. The detailed hyperparameters are listed in Table 10. Note that a significantly higher learning rate is used for the MLP head (5e-2) compared to the backbone (2e-5), for accelerating the convergence of the classification boundary while preserving the pre-trained linguistic features.

E.3 Prompt Templates

To enable the model to handle both concrete morpheme combinations and abstract semantic compositions, four distinct prompt templates 'during training. Below we present the specific instructions used. Note that `{char}`, `{pos}`, `{sense}`, and `{r}` serve as placeholders for the morpheme (character form), morpheme POS tag, morpheme sense, and word-formation, respectively.

Template 1: Full Word Input

[Chinese Prompt]
我将给你一个汉语二字词，你需要判断它是否符合汉语构词规则，是一个合理的二字词。该词为 **{char1}{char2}**，其中第一个字的词性为 **{pos1}**，意义为 **{sense1}**，第二个字的词性为 **{pos2}**，意义为 **{sense2}**。该词的构词结构为 **{r}** 结构。你的判断是：

[English Translation]
I will give you a Chinese disyllabic word. You need to judge whether it conforms to Chinese morphological rules and is a valid word. The word is **{char1}{char2}**. The first character has POS **{pos1}** and meaning **{sense1}**; the second character has POS **{pos2}** and meaning **{sense2}**. The word-formation is **{r}**. Your judgment is:

Template 2: First Morpheme Masked

[Chinese Prompt]
我将给你一个不完整的汉语二字词，你需要判断它是否符合汉语构词规则，是一个合理的二字词。该词为 **X{char2}**，其中第一个字的词性为 **{pos1}**，意义为 **{sense1}**，第二个字的词性为 **{pos2}**，意义为

Sample Type	Strategy Description	Example (Based on "季春")
Positive Samples *(Valid Semantics)*		
Existing Word	Original dictionary entry	**季春** (Last Spring Month)
Category Subst.	Replace morpheme with Hypernym	**季 <Season>** (Last <Season>)
Sibling Subst.	Replace with semantic sibling	**季夏** (Last Summer Month)
Negative Samples *(Invalid Semantics)*		
Random	Random morpheme replacement	**季三** (Last Three)
POS-constrained	Same POS, incompatible meaning	**季水** (Last Water)
High-Frequency	Common character replacement	**季我** (Last Me)

Table 9: Examples of the data augmentation strategy. The source word is **季春**, a *Modifier-Head* noun meaning *"The spring month"*. Here, "季" means "The third month of a season", and "春" means "Spring". Note that while we illustrate substitution on the tail morpheme ("春") here for clarity, our actual pipeline performs substitution on both head and tail morphemes stochastically.

Hyperparameter	Value
Model Architecture	Qwen3-Embedding-4B + MLP
Backbone Learning Rate	2e-5
MLP Head Learning Rate	5e-2
Loss Function	$\mathcal{L}_{\text{CE}} + \alpha \cdot \mathcal{L}_{\text{SupCon}}$
SupCon Loss Weight (α)	0.5
Train Batch Size	16
Gradient Accumulation Steps	4
Number of Epochs	10
Optimizer	AdamW
Precision	BF16
DeepSpeed Stage	2
SupCon Temperature	0.07
Validation Metric (Accuracy)	**95.23%**
Validation Metric (F1-Score)	**84.16%**

Table 10: Hyperparameters and final performance metrics for fine-tuning the semantic filtering model.

{sense2}。该词的构词结构为 **{r}** 结构。你的判断是：

[English Translation]
I will give you an incomplete Chinese disyllabic word. You need to judge whether it conforms to Chinese morphological rules and is a valid word. The word is **X{char2}**. The first character has POS **{pos1}** and meaning **{sense1}**; the second character has POS **{pos2}** and meaning **{sense2}**. The word-formation is **{r}**. Your judgment is:

Template 3: Second Morpheme Masked

[Chinese Prompt]
我将给你一个不完整的汉语二字词，你需要判断它是否符合汉语构词规则，是一个合理的二字词。该词为 **{char1}X**，其中第一个字的词性为 **{pos1}**，意义为 **{sense1}**，第二个字的词性为 **{pos2}**，意义为 **{sense2}**。该词的构词结构为 **{r}** 结构。你的判断是：

[English Translation]
I will give you an incomplete Chinese disyllabic word. You need to judge whether it conforms to Chinese morphological rules and is a valid word. The word is **{char1}X**. The first character has POS **{pos1}** and meaning **{sense1}**; the second character has POS **{pos2}** and meaning **{sense2}**. The word-formation is **{r}**. Your judgment is:

Template 4: Fully Abstract / Both Masked

[Chinese Prompt]
我将给你一个不完整的汉语二字词，你需要判断它是否符合汉语构词规则，是一个合理的二字词。该词中，第一个字的词性为 **{pos1}**，意义为 **{sense1}**，第二个字的词性为 **{pos2}**，意义为 **{sense2}**。该词的构词结构为 **{r}** 结构。你的判断是：

[English Translation]
I will give you an incomplete Chinese disyllabic word. You need to judge whether it conforms to Chinese morphological rules and is a valid word. In this word, the first character has POS **{pos1}** and meaning **{sense1}**; the second character has POS **{pos2}** and meaning **{sense2}**. The word-formation structure is **{r}**. Your judgment is:

F Prompt for Verification via Closed-Source LLMs

In the Verification stage, closed-source LLMs are employed to verify the semantic validity of candidate neologisms. The models operate under a specific expert persona and set of criteria to ensure rigorous judgment. Below, the exact system prompt and user prompt template are provided.

System Prompt

[Chinese Original]
你是一位严谨的汉语词汇语义专家，需要依据给定的构词信息判断一个候选二字词是否能在汉语中成立。请严格结合语法、语义与语感三方面，确保词性搭配、语义逻辑与常用语感一致，并注意读音不要与常见词产生歧义。

[English Translation]
You are a rigorous Chinese lexico-semantic expert, tasked with judging whether a candidate disyllabic word is viable in Chinese based on the provided morphological information. Please strictly integrate morphological, semantics, and linguistic intuition, ensuring appropriate POS collocation, semantic coherence, and consistency with common linguistic intuition. Also, pay attention to avoiding phonetic ambiguity with common existing words.

User Prompt Template

[Chinese Original]
请审查下述候选新词并仅返回 JSON。判断要求：
1. 候选词由【首字】和【尾字】按给定语法关系组合，需符合汉语构词习惯；
2. 若组合能表达自洽、易懂且可在真实语境中使用的概念，则判定为 YES；
3. 若语义牵强、搭配不当或违背语法语义规则，则判定为 NO；
4. 严格依据以下三条判断标准：
(1) 语法上合理——满足给定语法关系下的构词规则，词性搭配恰当；
(2) 语义上合理——能表达自洽、易懂且与两字释义相关的意义，避免牵强附会；
(3) 语感可接受——常用语感/读音自然，不与常见常用词读音或意义形成混淆。
5. 请在 reason 中用不超过 40 个汉字解释理由。
候选信息：
- 首字：**{head_char}**（**{head_pos}**，释义：**{head_sense}**）
- 尾字：**{tail_char}**（**{tail_pos}**，释义：**{tail_sense}**）
- 语法关系：**{relation}**
- 几个可能的语义组合示例：
* 将“首字”视为主要语义成分，尾字作搭配补充；
* 若语法关系为定中/主谓/状中，请按常见构词规则分析其自然度。
输出格式（必须是单个 JSON 对象，不得包含额外文字）：
{
“word”: “**{candidate_word}**”,
“judgement”: “YES 或 NO”,
“reason”: “简短理由”
}

[English Translation]
Please review the following candidate neologism and return ONLY a JSON object. Judgment Requirements:
1. The candidate is formed by combining [Head Char] and [Tail Char] according to the given word-formation, and must conform to Chinese morphological habits;
2. Judge as YES if the combination expresses a self-consistent, intelligible concept usable in real contexts;
3. Judge as NO if the meaning is far-fetched, the collocation is inappropriate, or it violates common morphological or semantic rule;
4. Strictly adhere to the following three criteria:
(1) Morphologically Reasonable: Meets formation rules under the relation; POS tagging is appropriate;
(2) Semantically Reasonable: Expresses a coherent meaning related to the characters, avoiding forced interpretations;
(3) Acceptable Intuition: Sounds natural in common usage/pronunciation, avoiding confusion with existing common words.
5. Please explain the reason in the ‘reason‘ field using no more than 40 Chinese characters.
Candidate Information:
- Head Char: **{head_char}** (**{head_pos}**, Sense: **{head_sense}**)
- Tail Char: **{tail_char}** (**{tail_pos}**, Sense: **{tail_sense}**)
- Relation: **{relation}**
- Examples of possible semantic combinations:
* View “Head Char” as the primary semantic component, with “Tail Char” as a supplement;
* If the relation is *Modifier-Head* / *Subject-Predicate* / *Adverbial-Head*, analyze naturalness based on common rules.
Output Format (Must be a single JSON object, no extra text):
{
“word”: “**{candidate_word}**”,
“judgement”: “YES or NO”,
“reason”: “Brief reason”
}

G Detailed Statistics of the Sets in Evaluation

This section provides a detailed breakdown of the sets in the evaluation section (Section 4). Table 11 presents the total number of candidate neologisms and their distribution across different word-formations for both the POS and NEG sets. For comparison, the table also includes the distribution of these word-formations in the existing Chinese lexicon, as derived from the COOL Dataset (see Table 5). The word-formations are ordered based on their prevalence in the POS set.

Analysis of Morphological Distribution. The comparative statistics in Table 11 offer two key observations regarding the generative behavior of the pipeline:

Distributional Alignment and Broad Coverage: The Positive set broadly mirrors the word-formation distribution of the Existing Lexicon, confirming that the generated neologisms align with the natural morphological patterns of Chinese. Beyond statistical alignment, this distribution notably demonstrates the pipeline’s capability to identify reasonable neologisms across a diverse range of word-formations. From dominant types like *Modifier-Head* and *Verb-Object* to other specific patterns, the pipeline consistently yields acceptable candidates, ensuring comprehensive coverage of the latent lexical space. In contrast, the Negative set shows a chaotic distribution, with a severe under-representation of canonical forms (e.g., *Modifier-Head* drops to 8.69%) and an over-abundance of typically strictly constrained struc-

Word-Formation	POS Set		NEG Set		Existing Lexicon
	Count	%	Count	%	%
定中 (Modifier-Head)	4,995	26.77%	51,292	8.69%	38.62%
述宾 (Verb-Object)	3,398	18.21%	44,163	7.49%	16.44%
联合 (Parallel)	3,145	16.85%	36,227	6.14%	22.87%
状中 (Adverb-Verb)	1,966	10.53%	71,602	12.14%	8.45%
方位 (Entity-Position)	1,616	8.66%	52,270	8.86%	0.37%
连谓 (Verb-Consequence)	1,039	5.57%	57,231	9.70%	3.43%
主谓 (Subject-Predicate)	754	4.04%	59,135	10.02%	1.06%
述补 (Verb-Complement)	572	3.07%	44,433	7.53%	1.28%
重叠 (Overlapping)	538	2.88%	12,817	2.17%	0.59%
后缀 (Suffixation)	323	1.73%	21,266	3.60%	2.70%
名量 (Noun-Quantifier)	255	1.37%	23,677	4.01%	0.13%
数量 (Number-Quantifier)	30	0.16%	23,290	3.95%	0.10%
介宾 (Preposition-Object)	18	0.10%	33,362	5.65%	0.31%
复量 (Complex-Quantifier)	12	0.06%	20,821	3.53%	0.03%
前缀 (Prefixation)	1	<0.01%	38,425	6.51%	0.10%
Total	**18,662**	**100.00%**	**590,011**	**100.00%**	—

Table 11: Comparative Distribution of Word-Formations. The table compares the prevalence of each structure across the generated Positive (POS) and Negative (NEG) sets, and the existing lexicon.

tures like *Subject-Predicate* (10.02%).

Exploration of Latent Spaces: While generally aligned, the Positive set diverges in specific categories such as *Entity-Position* (8.66% vs. 0.37%) and *Subject-Predicate* (4.04% vs. 1.06%). The higher prevalence of these types suggests that the pipeline does not merely replicate the existing dictionary distribution. Instead, it actively explores the "productive but sparse" regions of the lexical space, identifying valid potential words in categories that are permissible in morphology but less saturated in the current lexicon.

H Annotation Guidelines for Human Judgment

This appendix provides the detailed instructions given to the expert annotators for evaluating the real-world acceptability of generated disyllabic neologisms.

H.1 Core Objective of the Task

The task asks annotators to judge whether a given disyllabic combination has the **intrinsic potential** to become a Chinese neologism. This judgment focuses on the combination itself: whether it is morphologically reasonable, semantically feasible, and intuitively acceptable as a possible Chinese word. LLM-generated example sentences are provided only as contextual references, not as the primary object of evaluation. Therefore, if an example sentence is poorly constructed or logically flawed but the target combination itself remains valid, annotators should ignore the sentence-level defect and still label the word as 1 (acceptable).

H.2 Definition of Annotation Labels

Annotators are asked to make a binary decision based on the following definitions:

Label	Core Judgment Logic
1 (Acceptable/Valid)	The combination is morphologically reasonable and semantically self-consistent. Furthermore, the intuitive linguistic sense derived from the literal characters (expected POS) aligns with the given definition's POS.
0 (Unacceptable/Invalid)	The combination is morphologically unreasonable, semantically conflicting (illogical), or problematic in terms of linguistic intuition.

Table 12: Binary Annotation Labels for Evaluating Neologism Potential. The table outlines the definitions and core judgment logic for the "Acceptable" (1) and "Unacceptable" (0) labels used in the manual annotation task.

H.3 Core Criteria of Evaluation (Layered Evaluation)

Annotators should evaluate each candidate in the following order: **L1 (Morphology)**, **L2 (Semantics)**, and **L3 (Pragmatics)**. The layers are ordered from structural validity to holistic acceptability. A candidate that clearly fails an earlier layer should normally be labeled unacceptable, even if a later interpretation can be imagined.

H.3.1 L1: Layer of Morphological Validation

Basis. This layer determines whether the two characters are combined in a way that conforms to Chinese word-formation rules. Annotators should first check whether the proposed structure is morphologically valid, such as *Modifier-Head* for Noun+Noun combinations or *Verb-Object* for Verb+Noun combinations. If the structure is incompatible with ordinary Chinese morphology, the candidate should be rejected at this layer. For example, an Adverb+Noun combination is usually unacceptable unless it has a clear idiomatic basis.

H.3.2 L2: Layer of Semantic Compatibility

Basis. This layer asks whether the two morphemes can form a logically coherent meaning. Annotators should not judge whether the supplied definition is elegantly written; the central question is whether the meanings of the two morphemes are compatible when combined. For instance, "报电" (bào diàn, report–telegram) is acceptable because reporting and telegrams can be naturally associated. By contrast, "红声" (hóng shēng, red–sound) is normally unacceptable because color does not modify sound in ordinary usage, unless a very specific synesthetic context is intended.

H.3.3 L3: Layer of Pragmatic Acceptability

This layer assesses whether the candidate is acceptable as a usable word from the perspective of native-speaker intuition. It covers the alignment between morphological intuition and definition, phonetic naturalness, analogy with established patterns, and overall communicative acceptability.

Alignment of POS Intuition The intuitive POS of the candidate should align with the POS implied by its definition. A mismatch signals a pragmatic conflict. For example, "吃书" (eat book) is intuitively a *Verb-Object* expression. If it is defined as a noun such as "a type of person who likes books", the conversion is unnatural in Chinese without an added suffix such as "者", and the candidate should be labeled unacceptable. If it is defined as a verb such as "to read rapidly", the definition aligns with the candidate's intuitive structure and can be accepted.

Phonetics and Blocking The candidate should sound natural and should not be strongly blocked by an identical-sounding common word. Annotators should consider whether the pronunciation is awkward, for example because of difficult tone stacking, and whether it would be confused with a highly frequent existing expression.

Analogy with Established Pattern Plausibility is strengthened when the candidate follows a productive and familiar word-formation template. For example, **"搜商"** (search quotient) is plausible because it follows the established "X-商" pattern, as in "智商" (IQ) and "情商" (EQ). In contrast, an obscure combination without a clear analogical precedent should be treated with more caution.

Subjective Acceptability As a final holistic check, annotators should decide whether the candidate could be accepted by a native speaker in communication. Acceptable candidates read smoothly, are easy to understand, and feel like useful or natural additions to the lexicon. Unacceptable candidates read awkwardly, feel unnecessary because an existing word is clearly better, or sound strange, uncouth, or stiff.

H.4 Case Study

Example 1: Positive Case. **报电** (bào diàn) is a *Parallel* formation composed of 报 (report/news) and 电 (telegram). The definition, "written reports sent via telegram", is acceptable because the two morphemes are semantically compatible: reporting and telegrams can naturally refer to a specific message carrier. Its POS intuition is also compatible with the definition, since "报电" can be read either as a verbal action or as a nominal message. **Result: Acceptable.**

Example 2: Conflict of POS Intuition. **打桌** (dǎ zhuō) is a *Verb-Object* formation composed of 打 (make/build) and 桌 (table). Although "make" and "table" are semantically compatible, the form is intuitively understood as a verb-object action, namely "to make a table". The supplied definition treats it as a noun meaning "a tool specifically used for making tables", which forces an unnatural category shift in Chinese. **Result: Unacceptable.**

Example 3: Semantic Incompatibility. **冷跑** (lěng pǎo) is analyzed as a *Modifier-Head* formation composed of 冷 (cold) and 跑 (run), with the intended meaning of running training in cold weather. Although the LLM definition is understandable, adjectives such as "冷" more naturally modify nouns or states, as in "冷饮" (cold drink) or "冷清" (desolate). The collocation "冷跑" therefore feels stiff and less natural than alternatives

such as “冬跑” (winter run) or “寒跑” (cold run). **Result: Unacceptable.**

Example 4: Bad Sentence but Good Word. **微感** (wēi gǎn) is a *Modifier-Head* formation composed of 微 (tiny/slight) and 感 (feeling/touched), with the intended meaning of a slight feeling or being slightly moved. The generated sentence “这个苹果非常好吃，我有**微感**。” is poor and nonsensical, but the word itself remains plausible: “微” and “感” combine smoothly, comparable to expressions such as “微痛” (slight pain) or “微醺” (slightly tipsy). Annotators should therefore ignore the defective example sentence and judge the candidate itself. **Result: Acceptable.**

H.5 Supplementary Notes

The Distinction between Words and Phrases For computational convenience, the strict definitions distinguishing between words and phrases is abandoned here (Dong, 2005). Instead, all valid forms created by existing morphological patterns are considered as words. (For example, while we subjectively feel “第一” (first) and “第二” (second) are more like words than “第九” (ninth) or “第二十” (twentieth), we acknowledge here that they are all words, even if they are not individually listed in the lexicon).

Archaic or Dialect Words If a word is archaic or dialectal and uncommon in general Chinese contexts, but it conforms to morphology and can be understood and accepted by speakers, it can be viewed as a word with “potential for revival/borrowing” and labeled as Acceptable.

I Calculation Details for PPL

In this section, we outline the mathematical formulations and implementation details for calculating PPL scores for both model architectures discussed in Section 4. Let S denote a context sentence, and $w = (c_1, c_2)$ denote the disyllabic target word within that sentence. It is worth noting that for the models used in our evaluation, each Chinese character is generally treated as a single token.

I.1 Bidirectional Masked Models

For Masked Language Models (e.g., BERT), Pseudo-Log-Likelihood (PLL) scores are employed. Since these models are trained to predict masked tokens given a bidirectional context, the characters of the target word are masked one by one.

Let $S_{\setminus c_i}$ represent the sentence where only character c_i is masked. The score for the word w is calculated as:

$$\text{PPL}_{\text{MLM}}(w) = \exp\left(-\frac{1}{2}\sum_{i=1}^{2}\log P_\theta(c_i \mid S_{\setminus c_i})\right) \tag{2}$$

where P_θ is the probability distribution given by the model parameters θ. This metric reflects how predictable each character is given the surrounding context and the other character in the word.

I.2 Causal Autoregressive Models

For Causal Models (e.g., Llama, Qwen), which generate text sequentially from left to right, direct access to future context is not possible when computing probabilities for a word in the middle of a sentence. To address this, the evaluation is adapted into a prompted generation task.

I.2.1 Prompt Construction

A set of prompt templates $\mathcal{T}$ is defined. For a sentence S containing the target word w, w is first replaced with a placeholder “~” (a wave dash) to create a masked context S_{masked}. This context S_{masked} is then concatenated with a template $t \in \mathcal{T}$ to form the model input prefix.

Three Chinese templates are utilized to query the model:

- Template 1: "~是：" (is:)
- Template 2: "这里的~指的是：" (The ~ here refers to:)
- Template 3: "句中的~表示：" (The ~ in the sentence indicates:)

I.2.2 Perplexity Calculation

For a given template t_i, the input prefix is $x = [S_{\text{masked}}; t_i]$. The model is tasked with generating the characters c_1 and c_2 immediately following this prefix. The score for a single template is:

$$\text{PPL}(w, t_i) = \exp\left(-\frac{1}{2}\Big(\log P_\theta(c_1 \mid x) + \log P_\theta(c_2 \mid x, c_1)\Big)\right) \tag{3}$$

The final PPL for the word w is the average score across all three templates in $\mathcal{T}$ to reduce variance caused by prompt sensitivity:

$$\text{PPL}_{\text{Causal}}(w) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\text{PPL}(w, t) \tag{4}$$

J Zero-Shot and Attested Candidate Analysis

We further separate generated POS candidates by corpus attestation to examine whether the framework mainly reproduces words already seen in text or explores unattested lexical possibilities. Using an 11B-token Chinese corpus as a posterior attestation source, we define C^+ as candidates with at least one corpus occurrence and C^- as candidates with zero corpus occurrence. Among 15,793 POS candidates, 14,841 candidates (94.0%) are unattested (C^-), while 952 candidates (6.0%) are attested (C^+). The zero-shot POS candidates still achieve lower PPL than the NEG set (7.29 vs. 8.49, Mann–Whitney U test, $p \approx 0$), indicating that the generated candidates are not explained merely by corpus memorization. The attested POS subset has a slightly lower PPL of 7.07, as expected for forms already observed in natural text.

Group	Count	Share	Avg. PPL
POS-C^- (zero-shot)	14,841	94.0%	7.29
POS-C^+ (attested)	952	6.0%	7.07
NEG reference	–	–	8.49

Table 13: Zero-shot and attested analysis of POS candidates using an 11B-token Chinese corpus as posterior evidence. Lower PPL for POS-C^- than NEG indicates that unattested candidates remain contextually plausible.

K Bootstrap Analysis of Human Judgment

To assess the robustness of the human judgment results, we conduct bootstrap resampling with $N = 1000$ iterations using stratified sampling with replacement over the 400 annotated samples (200 POS and 200 NEG). The 95% confidence intervals for POS and NEG do not overlap, and the estimated effect size is large (Cohen's $d = 14.95$). This confirms that the gap between POS and NEG is stable under resampling.

Set	Mean Acc.	95% CI	Half-width
POS	71.9%	[65.33%, 77.89%]	6.3%
NEG	24.8%	[18.99%, 31.00%]	6.0%

Table 14: Bootstrap analysis of human judgment with 1,000 stratified resampling iterations. The POS and NEG confidence intervals do not overlap, with Cohen's $d = 14.95$.

L Attestation Effect on Human Judgment

We also examine whether human acceptance is driven by corpus attestation. In the POS sample, only 10 of 200 candidates are attested in the corpus; the remaining 190 unattested POS candidates still achieve a 72.1% acceptance rate (137/190 accepted). In the NEG sample, 199 of 200 candidates are unattested, and the NEG class remains much less acceptable than POS overall (25.0% vs. 71.5%, $p < 10^{-20}$). These results suggest that human validity judgments are driven primarily by morphological and semantic plausibility rather than by corpus presence.

Group	Count	Accepted	Acc. Rate
POS-C^+	10	–	–
POS-C^-	190	137	72.1%
NEG-C^-	199	–	–
POS overall	200	–	71.5%
NEG overall	200	–	25.0%

Table 15: Effect of corpus attestation on human judgment. The rebuttal analysis reports detailed accepted counts for POS-C^- and overall POS/NEG acceptance rates; unavailable cell values are left blank rather than inferred.

M Detailed PPL Analysis by Word-Formation

In this section, a granular breakdown of the PPL scores across different word-formations is presented. This analysis allows for the investigation of whether the quality of generated neologisms varies depending on their word-formations.

To ensure statistical reliability, word-formations with insufficient sample sizes are excluded ($N < 50$ in the POS set), such as *Prefixation* (前缀), *Preposition-Object* (介宾), *Number-Quantifier* (数量), and *Complex-Quantifier* (复量). The filtered results for the major word-formations are summarized in Table 16.

The breakdown reveals consistent trends that reinforce our main findings:

Superiority of Consistency. Across all 11 analyzed word-formations, the POS set consistently exhibits lower PPL scores compared to the NEG set. This indicates that our pipeline's effectiveness is robust and not biased towards specific word-formation patterns;

Word-Formation	Mean Average PPL		
	EXIST	POS	NEG
主谓 (Subject-Predicate)	5.46	7.00	8.25
定中 (Modifier-Head)	5.38	7.03	7.97
状中 (Adverb-Verb)	5.69	7.26	8.24
联合 (Parallel)	5.78	7.32	8.09
述宾 (Verb-Object)	5.78	7.41	8.10
连谓 (Verb-Consequence)	5.66	7.51	8.34
述补 (Verb-Complement)	5.59	7.14	8.39
重叠 (Overlapping)	5.47	6.67	7.60
后缀 (Suffixation)	6.17	7.22	8.61
方位 (Entity-Position)	5.91	7.17	8.78
名量 (Noun-Quantifier)	4.83	7.30	8.85

Table 16: Breakdown of PPL scores by major word-formations. The POS set consistently achieves lower PPL (higher naturalness) than the NEG set across all word-formations.

Structures with Generation Difficulty. Verb-centric word-formations, such as *Verb-Object* (述宾) and *Verb-Consequence* (连谓), tend to have slightly higher PPL scores in the POS set (7.41 and 7.51, respectively). This suggests that generating valid neologisms involving transitive relations or temporal sequences is inherently more challenging than generating static nominal structures like *Modifier-Head* (定中, 7.03);

Categories with High Naturalness. The *Overlapping* (重叠) word-formation achieves the lowest PPL (6.67) in the POS set. This implies that reduplication is a highly productive and permissive mechanism in Chinese morphology. Additionally, word-formations like *Entity-Position* (方位) and *Noun-Quantifier* (名量) show a substantial gap between their POS scores (7.17, 7.30) and NEG scores (8.78, 8.85). This significant contrast highlights the model's capability to capture the strict semantic constraints inherent in these structures, effectively distinguishing valid concepts from nonsensical ones.

N Error Analysis

We conduct a qualitative error analysis on cases where the automatic pipeline and human judgments disagree. False positives are POS candidates rejected by the majority of annotators, while false negatives are NEG candidates accepted by the majority of annotators. Table 18 summarizes representative examples. The main failure modes are: (i) semantically interpretable but unidiomatic combinations, (ii) mismatches between the intended definition and the word-formation intuition, (iii) blocking by more natural existing expressions, and (iv) over-filtering of structurally valid candidates.

O Case Study of Generated Neologisms

To provide a qualitative understanding of our generation and filtering process, a comparative case study is presented, with examples selected from the POS and NEG sets. Table 17 lists three representative neologisms from each set, along with their morphological components, definitions, and an analysis of their validity.

Word	Components & Structure	Definition & Example	Analysis
		Positive Examples (Accepted)	
兵势 (bīng shì)	**Head:** 兵 (Army) **Tail:** 势 (Momentum) **Word-Formation:** *Modifier-Head*	**Def:** Military power and its influence. **Ex:** 这次演习旨在展示该国的**兵势**。(The drill aims to show the country's *military-momentum*.)	**Valid.** Strong semantic composition ("Army's momentum"). **Avg PPL: 6.69**
访诊 (fǎng zhěn)	**Head:** 访 (Visit) **Tail:** 诊 (Diagnose) **Word-Formation:** *Adverb-Verb*	**Def:** To visit a patient for diagnosis (house call). **Ex:** 医生决定上门**访诊**为他检查。(The doctor decided to *visit-diagnose* him at home.)	**Valid.** Useful medical term filling a lexical gap. **Avg PPL: 6.43**
练操 (liàn cāo)	**Head:** 练 (Practice) **Tail:** 操 (Drill) **Word-Formation:** *Verb-Sequence*	**Def:** To practice and drill skills repeatedly. **Ex:** 掌握技能离不开反复的**练操**。(Mastering skills requires repeated *practice-drill*.)	**Valid.** Natural parallel combination of synonyms. **Avg PPL: 6.83**
		Negative Examples (Rejected)	
子麦 (zǐ mài)	**Head:** 子 (Suffix) **Tail:** 麦 (Wheat) **Word-Formation:** *Subject-Predicate*	**Def:** Small wheat. **Ex:** 我国北方主要种植**子麦**。(Northern China mainly grows *suffix-wheat*.)	**Invalid Structure.** "子" is typically a suffix (e.g., 麦子). Using it as *Subject-Predicate* is ungrammatical. **Avg PPL: 7.73**
流眠 (liú mián)	**Head:** 流 (Flow) **Tail:** 眠 (Sleep) **Word-Formation:** *Adverb-Verb*	**Def:** Sleep is unstable or flowing. **Ex:** 她整夜都睡得**流眠**。(She slept *flowingly* all night.)	**Semantic Conflict.** "Flow" (dynamic) contradicts "Sleep" (static). **Avg PPL: 8.63**
高卖 (gāo mài)	**Head:** 高 (High) **Tail:** 卖 (Sell) **Word-Formation:** *Adverb-Verb*	**Def:** To sell at a high price. **Ex:** 这家商店总是把商品**高卖**。(This shop always *high-sells* goods.)	**Unidiomatic.** Morphologically correct but Chinese prefers "高价卖". **Avg PPL: 6.48** (Low PPL but rejected).

Table 17: Case study of generated neologisms. Positive examples demonstrate morphological, semantic and pragmatic validity, while negative examples exhibit morphological errors, semantic conflicts, or lack of idiomaticity. Average PPL scores are calculated across four models.

Type	Word	Formation	Definition	Error Analysis
False positives: POS candidates rejected by human annotators				
FP	功额	Noun-Quantifier	A quantitative standard for performance or task effects.	The intended meaning is understandable, but the form is blocked by more natural expressions such as *quota* or *performance target*. The noun-quantifier relation also feels forced because 额 already carries the sense of a fixed amount.
FP	作病	Verb-Object	To cause a disease or trigger symptoms.	The definition shifts from the morpheme 作 "occur/happen" toward a causative reading. Native intuition prefers 致病 or 发病, making the generated form semantically strained.
FP	补献	Adverb-Verb	To make an additional offering or donation to compensate for a shortfall.	The two morphemes are compatible, but 补 modifying 献 is pragmatically stiff. More conventional forms such as 补捐 or 追加捐赠 better express the intended meaning.
False negatives: NEG candidates accepted by human annotators				
FN	椅罩	Modifier-Head	A cover placed over a chair for protection or decoration.	This is a transparent and productive compound. The negative label is likely due to over-filtering rather than linguistic invalidity.
FN	抽校	Verb-Sequence	To extract a subset for checking or correction.	The form naturally denotes sampling followed by verification, a plausible technical verb in inspection or auditing contexts.
FN	战外	Entity-Position	The area or scope outside a war or conflict zone.	Although uncommon, the entity-position relation is clear and interpretable. The candidate illustrates a valid low-frequency spatial compound that the pipeline can underrank.

Table 18: Qualitative error analysis of disagreement cases between the automatic pipeline labels and majority human judgments. FP denotes POS candidates rejected by humans; FN denotes NEG candidates accepted by humans.