From 124 Million Tokens to 1,021 Neologisms: A Large-Scale Pipeline for Automatic Neologism Detection

Diego Rossini, Lonneke van der Plas
Università della Svizzera italiana (USI)
Lugano, Switzerland
{diego.rossini, lonneke.vanderplas}@usi.ch

Abstract

We present a scalable, modular pipeline for automatic neologism detection that combines rule-based filtering with LLM classification. The pipeline is grounded in two complementary word-formation frameworks, grammatical and extra-grammatical morphology, which jointly define the scope of what counts as a neologism and inform a four-class classification scheme (NEOLOGISM, ENTITY, FOREIGN, NONE). While designed to be modular and transferable at the architectural level, the pipeline is instantiated on 527 million English-language Reddit posts spanning 2005–2024. From this corpus, we extract 124.6 million unique tokens and reduce them by over 99.99% to yield 1,021 neologism candidates, a set small enough for manual expert verification. Multiple LLMs independently classify each candidate via majority vote, with a final verification step, revealing substantial cross-model disagreement and highlighting the challenge of operationalizing neologism detection at scale. Manual annotation of all 1,021 candidates confirms that 599 (58.7%) are genuine lexical innovations. The pipeline code, vocabulary compilation scripts, and the annotated candidate list are available at https://github.com/DiegoRossini/neologism-pipeline.

Keywords: neologism detection, lexical innovation, Reddit, large language models, rule-based filtering, computational neology

1. Introduction

Although the study of neologisms has deep roots in linguistics (Guilbert, 1975; Rey, 1976), their automatic detection is a comparatively recent task. Computational approaches only became feasible once large machine-readable corpora were available in the 1990s (Renouf, 1993; Cabré and de Yzaguirre, 1995). Since then, a number of web-based platforms have been developed for neologism identification (Kerremans et al., 2012; Cartier, 2017; Klosa-Kückelhaus and Lüngen, 2018). These systems typically depend on static exclusion dictionaries and language-specific resources, and require manual expert verification of their output (Brasolin et al., 2023; Cartier, 2017; Tomaszewska et al., 2025). The key challenge for any detection pipeline is therefore to reduce the candidate set to a size where such verification is feasible.

More recently, social media data has attracted increasing attention as a source for studying lexical innovation, given the volume, diversity, and informality of user-generated content (Grieve et al., 2018; Würschinger, 2021). However, the same characteristics that make these platforms attractive also pose a challenge for neologism detection: in a large corpus, the vast majority of tokens absent from standard dictionaries are not neologisms but typos, misspellings, concatenated strings, code fragments, or foreign-language material. In the dataset used in this study, 527 million English-language Reddit posts spanning 2005–2024 (Baumgartner et al., 2020; Watchful1, 2025), we extract 124.6 million unique tokens, which the pipeline reduces by over 99.99% to yield 1,021 neologism candidates. Classifying each of the 124.6 million unique tokens individually with an LLM would be computationally infeasible, which motivates a multi-stage filtering approach that progressively narrows the set before classification.

In this paper, we present a pipeline for large-scale neologism detection that combines deterministic rule-based filtering with LLM-based classification. The rule-based stages progressively reduce the candidate set; the LLM stage then classifies surviving candidates into four categories: ENTITY, NEOLOGISM, FOREIGN, or NONE. Multiple LLMs independently classify each candidate, and only those receiving a majority vote are retained. A final verification step can then confirm or discard the output of the preceding models. Our contributions are: (1) a scalable, modular pipeline for neologism detection from social media corpora, grounded in word-formation theory (§3), whose output illustrates a range of grammatical and extra-grammatical word-formation processes (§7.1); (2) a comparative evaluation of multiple LLMs on a four-class neologism classification task; and (3) a detailed manual analysis of all 1,021 pipeline output candidates, including gold annotation, error analysis by category (§7.2), and classification of detected neologisms along word-formation processes (§7.1).[1]

[1]The pipeline code, vocabulary compilation scripts, and the annotated candidate list are available at https://github.com/DiegoRossini/neologism-pipeline.

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 1–15
16 MAY 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Related Work

The dominant paradigm for automatic neologism detection remains the *exclusion dictionary method*: a token is flagged as a candidate neologism if it does not appear in one or more reference lexicons (Renouf, 1993; Cabré and de Yzaguirre, 1995). This principle underpins the major detection platforms developed over the past two decades, including the NeoCrawler (Kerremans et al., 2012, 2018), which monitored English-language websites for previously unattested forms, Néoveille (Cartier, 2017), which adopted a similar architecture for multiple languages, and the IDS Neologismenwörterbuch (Klosa-Kückelhaus and Lüngen, 2018), a continuously updated German neologism dictionary backed by corpus monitoring. The NeoCrawler was formally decommissioned in 2020 (Q. Würschinger, personal communication, 2025), illustrating the fragility of long-term tool availability. All three systems depend on language-specific resources that limit portability across languages and corpora. Our pipeline adopts the same foundational exclusion principle but separates language-specific resources from the architectural design: the sequence of filtering stages is pre-determined, while the resources they operate on (reference vocabularies, phonotactic rules, frequency dictionaries) should be substituted or adapted for each target language.

Beyond pure exclusion lookup, Falk et al. (2014) trained an SVM on French newspaper candidates using form-related, morpho-lexical, and thematic features, demonstrating the value of semantic context for neologism classification. Our pipeline follows a similar two-stage logic, but delegates the classification step to prompted LLMs rather than to feature-based classifiers.

As corpora drawn from social media have grown in size, so has the difficulty of the candidate extraction step itself. Grieve et al. (2018) identified 54 emerging words from 8.9 billion tokens of geolocated American Twitter data by tracking frequency increases over time and filtering manually. Mahler (2020) applied a comparable frequency-based methodology to Reddit, and Würschinger (2021) demonstrated how network metrics capture properties of lexical innovation that frequency measures alone cannot reveal. Brasolin et al. (2023) and Spina et al. (2024) extracted candidates from millions of geolocated Italian tweets through exclusion filtering and manual distillation, identifying hundreds of unattested word forms. A recurring challenge across these studies is that the vast majority of tokens absent from standard dictionaries are not neologisms but typos, misspellings, code fragments, or foreign-language material.

The most directly comparable recent system is NeoN (Tomaszewska et al., 2025), a multi-layered pipeline for Polish that combines frequency analysis, structural constraints, reference corpus checking, and spelling error detection with an LLM-based final filter, demonstrating that LLMs can serve as effective precision boosters after rule-based prefiltering. Our work differs from NeoN in several respects: we ground the pipeline in two complementary word-formation frameworks (§3) that define the scope of each category; we adopt a four-class taxonomy rather than a binary filter; we use a multi-model majority-vote scheme with independent verification rather than a single LLM; and we provide a detailed manual analysis of all pipeline output, including error categorisation and classification by word-formation process.

3. Theoretical Foundations

Any neologism detection pipeline presupposes an operational definition of what counts as a neologism. This section presents the two word-formation frameworks that jointly inform the design of our classification scheme and, in particular, determine the scope of the NEOLOGISM label assigned by the LLM stage (§4).

3.1. Grammatical Word Formation

Štekauer's onomasiological theory (Štekauer, 1998; Štekauer, 2001; Štekauer, 2005) models word formation as a top-down, need-driven process: a speaker identifies a concept lacking a conventional expression and coins a new naming unit by selecting an onomasiological type — a structural pattern that maps conceptual content onto morphological form. The process is *grammatical* in the sense that, given a naming need, the resulting formation is constrained by the productive onomasiological types available in the language.

A central consequence of this framework concerns nonce-formations. Against the view that these are deviant, context-dependent, and inherently non-lexicalisable coinages (cf. Hohenhaus, 1998), Štekauer (2002) argues that nonce-formations are regular products of the Word-Formation Component, generated by the same productive rules as any other naming unit. What distinguishes them is not structural deviance but *lifecycle stage*: a nonce-formation is a neologism at the earliest point between coinage and dissemination, and whether it subsequently becomes institutionalised or falls out of use is an empirical matter that cannot be predicted at the time of coining. The notion of "nonce-formation" as a structurally distinct category thus collapses into a temporal label.

For the pipeline, this entails that tokens attested infrequently in the corpus cannot be excluded as

non-neologisms on formal grounds alone, since nonce-formations are structurally indistinguishable from formations that will eventually become established. The frequency threshold (§4.5) is accordingly designed to filter noise rather than to impose a lexicalisation requirement. However, Štekauer's model is explicitly limited to rule-governed formation: processes such as blending, clipping, and acronymy, whose output cannot be derived from productive onomasiological types, fall outside the Word-Formation Component and are relegated to the Lexicon (Štekauer, 2001). The following framework addresses precisely this gap.

3.2. Extra-Grammatical Word Formation

Mattiello (2013) addresses those processes that fall outside the scope of grammatical morphology: clippings, blends, acronyms, abbreviations, and other formations whose input does not allow prediction of a regular output through any rule-based model of word formation. Within the framework of Natural Morphology (Dressler, 2000), these are classified as *extra-grammatical*, distinct from both core grammatical morphology (rule-governed, productive) and marginal morphology (partially regular). Although traditionally marginalised for their irregularity and unpredictability, Mattiello (2013) demonstrates that extra-grammatical processes are productive in their own right, particularly in informal registers, and that they comply with criteria of well-formedness and contextual suitability. The driving mechanism is *analogy* rather than rule application (Mattiello, 2013, 2017; Arndt-Lappe, 2015): new formations arise by modelling on existing words, either individually (surface analogy) or through recurrent patterns (analogy via schema). Mattiello (2017), following Booij (2010) and Plag (1999), extends this mechanism to playful coinages such as *doggo* (← *dog*), previously excluded as *expressive morphology*, i.e. playful or affective modifications of existing words (Zwicky and Pullum, 1987). Moreover, the boundary between extra-grammatical and grammatical morphology is not fixed, as formations that originate as creative coinages can over time become regular and productive (Körtvélyessy et al., 2022, 2021).

For the pipeline, this entails that the NEOLOGISM class must be broad enough to encompass both grammatical and extra-grammatical formations. Taken together, the two frameworks define the theoretical scope of the positive labels used in this study: the four-class classification scheme presented in §4 operationalises this joint definition, with the NEOLOGISM and ENTITY classes capturing genuine lexical innovations and FOREIGN and NONE isolating non-neologistic material that rule-based filtering alone cannot eliminate.

4. Methodology

The pipeline is designed to be modular: the sequence of filtering stages is pre-determined, but the resources each stage operates on (reference vocabularies, phonotactic rules, frequency dictionaries) are language-specific and must be substituted for each target language. The instantiation described below targets English.

4.1. Tokenization

Raw texts are tokenized using a spaCy language model, with named entity recognition, dependency parsing, and lemmatization disabled for efficiency. The choice of model depends on the target language. A corpus-specific preprocessing step removes or replaces non-lexical content before tokenization: for social media corpora, this may include URLs, platform-specific references, user mentions, hashtags, emojis, and non-ASCII characters; for other corpus types, different noise patterns (e.g., markup tags, metadata fields) may require analogous treatment. Punctuation, stopwords, and whitespace tokens are discarded, and all remaining tokens are lowercased.

4.2. Vocabulary Filtering

Tokens present in a reference vocabulary compiled exclusively from sources predating a chosen cutoff date are filtered out on the assumption that they represent established lexical items rather than neologisms. The cutoff defines an observation window over which the lifecycle of detected neologisms can be tracked. The pipeline accepts one or more reference vocabularies; when multiple sources are available, combining them reduces the risk of false positives caused by gaps in any individual lexicon. The composition can be tailored to the target language and corpus: for social media data, it may include platform-specific vocabulary, crowdsourced slang dictionaries, and encyclopaedic entries alongside standard lexicons, while for more formal corpora, curated dictionaries and domain-specific terminologies may suffice. Any token found in the reference vocabulary is excluded from further processing.

4.3. Pattern-based Cleaning

Tokens surviving the vocabulary filter are subjected to pattern-based rules designed to remove noise that no dictionary would capture. These rules fall into two categories. The first is language-independent: tokens must be purely alphabetic and fall within a configurable length range, while those exhibiting excessive character repetition, low character entropy, or repeated character sequences are

discarded as likely keyboard spam or encoding artefacts. The second category is language-specific and must be adapted to the target language: this includes phonotactic constraints (e.g., implausible consonant or vowel clusters), expressive spelling variants (e.g., elongated interjections, laughter patterns), and corpus-specific noise patterns (e.g., placeholder or template text).

4.4. Typo and Concatenation Detection

Corpora, particularly those drawn from social media, frequently contain misspellings and tokens formed by words concatenated without spaces. Neither constitute lexical innovations, yet both survive vocabulary filtering because they do not match any reference entry. The pipeline applies SymSpell (Garbe, 2012), a symmetric delete spelling correction algorithm with support for multiple languages, to detect both cases against a reference frequency dictionary. A token is flagged as a typo if it falls within a configurable edit distance of a high-frequency entry in the dictionary, and as a concatenation if it can be segmented into two or more parts each appearing in the same dictionary. Minimum character length thresholds prevent spurious matches on short tokens, and a conservative maximum edit distance ensures that only tokens closely resembling high-frequency dictionary entries are flagged, so that morphologically complex forms such as compounds or blends are unlikely to be flagged; those that are can be recovered by the frequency-based reintegration mechanism (§4.5).

4.5. Frequency Threshold and Reintegration

Tokens previously excluded as typos or concatenations (§4.4) are reconsidered if they meet a configurable frequency threshold. If a token resembles a misspelling or a segmentable string yet recurs frequently in the corpus, it is unlikely to be an error, and its reintegration prevents genuine coinages from being prematurely discarded.

Candidates occurring fewer than the frequency threshold are excluded. While this introduces a tension with the theoretical position outlined in §3.1, where nonce-formations are treated as legitimate neologisms regardless of their frequency, the constraint is computational rather than theoretical: when too many candidates survive the rule-based filters, manual or LLM-based verification becomes infeasible. The threshold can be adjusted or omitted entirely depending on corpus size and available resources. In practice, nonce-formations attested below the threshold are lost; however, the threshold is not designed to impose a lexicalisation requirement but to separate deliberate, repeated use from accidental variation, since typos and random strings rarely recur at scale.

4.6. Foreign Language Detection

Depending on the target language, the corpus may contain substantial material from other languages that survives vocabulary filtering. The pipeline applies the Lingua language detector (Stahl, 2022) to flag and filter tokens identified as belonging to a language other than the target. A configurable confidence threshold controls how aggressively tokens are filtered. Tokens whose confidence score falls below the threshold, for instance due to mixed Tagalog–English morphology or orthographic overlap with the target language, are retained and delegated to the LLM classification stage, which includes a dedicated FOREIGN category. The stage does not distinguish foreign-language noise from loanwords entering the target language; this limitation is discussed in the Limitations section.

4.7. LLM Classification

Candidates surviving the filtering stages are classified into four categories using large language models. The taxonomy reflects the theoretical scope established in §3: NEOLOGISM (new words, slang, or words derived from proper nouns, encompassing both grammatical and extra-grammatical formations); ENTITY (proper nouns such as people, companies, brands, products, or places); FOREIGN (words from other languages that escaped the language detection stage, §4.6); and NONE (residual noise including usernames, typos, programming terms, and unclear cases).

Both NEOLOGISM and ENTITY constitute lexical innovations: in Štekauer's onomasiological framework, word-formation is a naming response to newly salient extra-linguistic referents, and proper nouns denoting emerging social or cultural entities qualify as newly established naming units. The two classes are kept separate for analytical purposes, facilitating comparison with standard NER categories in downstream applications. Classification is performed in two stages. First, multiple LLMs independently classify each candidate token. Labels are aggregated via majority vote: a token receives a label only if the majority of models agree; otherwise it is marked UNKNOWN. Second, an additional model verifies each label and produces the final output. The choice and number of models is configurable; using multiple architectures trained on different data reduces idiosyncratic misclassifications, while the independent verification step provides an additional quality control layer at minimal additional cost.

5. Experimental Setup

This section describes the instantiation of the pipeline for English-language neologism detection on Reddit data. All language-specific resources, parameters, and model choices reported below can be substituted for other languages or corpora.

5.1. Corpus

The corpus consists of Reddit submissions and comments spanning January 2005 to December 2024, drawn from the Pushshift archive (Baumgartner et al., 2020; Watchful1, 2025). After excluding deleted posts, removed content, and non-textual submissions, the dataset comprises approximately 527 million posts. Although the corpus is predominantly English, it contains multilingual content, most notably Taglish (Tagalog–English code-switching) in Filipino-oriented subreddits, as well as posts in Portuguese, Spanish, French, German, and other languages.

5.2. Tokenization

We use spaCy's `en_core_web_lg` model with named entity recognition, dependency parsing, and lemmatization disabled for efficiency. URLs, subreddit references (`r/\w+`), user mentions (`u/\w+`), and hashtags are replaced with placeholder tokens; emojis and non-ASCII characters are removed. Punctuation, stopwords, and whitespace tokens are discarded, and all remaining tokens are lowercased. The tokenization stage yields 124.6 million unique token types.

5.3. Reference Vocabularies

The reference vocabulary is compiled exclusively from pre-2015 sources, establishing a ten-year observation window (2015–2024) over which newly emerged tokens can be identified. The combined vocabulary comprises 16.3 million unique surface forms drawn from six sources (Table 1). Using multiple independently compiled resources reduces the risk of false positives caused by gaps in any individual lexicon: Reddit and Urban Dictionary (Urban Dictionary, 2025) cover informal register, Wikipedia titles (Wikimedia Foundation, 2015a) capture named entities and technical terminology, while WordNet (Princeton University, 2011), Wiktionary (Wikimedia Foundation, 2015b), and NoSlang (5.5K tokens, obtained with permission from the site owner) provide baseline lexical coverage.

5.4. Filtering Parameters

Pattern cleaning. Tokens must be purely alphabetic and between 3 and 20 characters in length. The English-specific rules filter tokens starting with double vowels (*aa*, *ee*, *ii*, *oo*, *uu*), implausible consonant clusters, expressive variants (*hahaha*, *yeaah*, *ughh*), repeated character sequences, and Lorem Ipsum placeholder words. Tokens exceeding six characters with two or fewer unique characters are discarded as keyboard spam. The full rule set is available in the project repository.

Typo and concatenation detection. SymSpell (Garbe, 2012) is configured with a maximum edit distance of 2 and a frequency dictionary compiled from Reddit pre-2015 token frequencies and WordNet. A token is flagged as a typo if its closest match in the dictionary has edit distance 1–2 and frequency above 100; minimum token length for typo checking is 5 characters. Concatenation detection applies word segmentation on tokens of at least 6 characters, flagging those that segment into two or more parts all present in the frequency dictionary. Genuine compounds flagged at this stage can be recovered by the reintegration mechanism described below.

Frequency threshold. The minimum occurrence threshold is set to 100 (§4.5). Tokens previously flagged as typos or concatenations are reintegrated if they meet this threshold.

Foreign language detection. The Lingua language detector (Stahl, 2022) is applied with a confidence threshold of 0.75 across 47 languages, removing 33,959 tokens (16.3% of the 208,932 candidates at that stage).

All thresholds reported above were set based on preliminary experimentation; a discussion of their limitations is provided in the Limitations section.

5.5. LLM Classification

The three open-source models—Qwen 2.5 72B, LLaMA 3.3 70B, and Mistral Large 2 123B—independently classify each candidate; labels are aggregated via majority vote (§4.7). Claude 4.5 Haiku serves as an independent verification source and does not participate in the vote.

All open-source models use few-shot prompting with eight labelled examples spanning the four classes and up to three contextual sentences per candidate, drawn from diverse subreddits. Claude 4.5 Haiku classifies tokens with the same contextual examples as the other models. The full prompt templates are provided in Appendix A.

5.6. Computational Setup

All experiments were run on a single multi-GPU server with 500 GB RAM and 4 GPUs (120 GB

Reference Vocabularies (all pre-2015)		
Source	**Tokens**	**Coverage**
Reddit pre-2015	10.5M	Informal, platform jargon
Wikipedia titles	4.4M	Entities, technical terms
Urban Dictionary	1.5M	Slang
Wiktionary	554K	Morphological variants
WordNet 3.1	147K	Core vocabulary
NoSlang	5.5K	Chat abbreviations
Total	**16.3M**	

Table 1: Reference vocabularies and primary coverage.

Stage	**Remaining**
Tokenization	124,593,754
Vocabulary lookup	113,909,871
Pattern cleaning	23,955,763
Concatenation detection	10,793,055
Typo detection	7,065,796
Freq. threshold + reintegration	208,932
Foreign language detection	174,973
Majority vote (NEOLOGISM)	10,499
Haiku verification	1,021

Table 2: Filtering cascade: candidates remaining after each stage.

each). The open-source models were sharded across all four GPUs in bfloat16 precision. Claude 4.5 Haiku was accessed via the Anthropic Batch API.

On the described hardware, the ideal critical path is approximately 50–65 hours (∼2–3 days): tokenization of 527 million posts accounts for 18–24 hours, vocabulary filtering and context retrieval for ∼9 hours, and sequential LLM inference over three models for 22–30 hours (40–50% of total compute). Running the three models in parallel on separate nodes would reduce the total to ∼38–49 hours.

6. Results

6.1. Filtering Cascade

Table 2 reports the number of candidate tokens surviving each pipeline stage. The rule-based stages reduce the initial 124.6 million unique tokens by 99.86%, yielding 174,973 candidates for LLM classification. The most aggressive single stage is pattern cleaning, which removes 90 million tokens (72.2% of the input at that point), followed by concatenation detection (13.2 million concatenated tokens) and vocabulary lookup (10.7 million known words). The frequency threshold eliminates a further 6.9 million low-frequency tokens. Of the tokens previously excluded as typos or concatenations, 118,544 meet the frequency threshold and are reintegrated into the candidate pool (§4.5).

6.2. LLM Classification and Inter-Model Agreement

The three open-source models independently classified all 174,973 tokens. Table 3 reports their label distributions, revealing substantial cross-model disagreement. LLaMA is the most aggressive NEOLOGISM predictor (12.2%, nearly double the other two models), while Mistral is the most conservative overall, assigning NONE to 59.9% of tokens. Qwen detects the most foreign-language material (22.2%).

Unanimous agreement across all three models is reached for only 45.8% of tokens (80,220); 48.4% are decided by a 2-out-of-3 majority, and 5.8% (10,134) result in three-way ties, conservatively resolved to NONE. These complementary biases validate the ensemble design: no single model would achieve the same coverage.

The majority vote produces 10,499 NEOLOGISM candidates (6.0%), 47,276 ENTITY (27.0%), 33,159 FOREIGN (19.0%), and 84,039 NONE (48.0%).

6.3. Haiku Verification

Claude 4.5 Haiku independently classified the same tokens with the same contextual examples. Applied as a verification filter to the 10,499 majority-vote NEOLOGISM candidates, Haiku confirmed 897 as NEOLOGISM (8.5%), relabeled 124 as ENTITY (1.2%), and rejected 9,478 to NONE (90.3%). This high rejection rate is driven primarily by model conservatism rather than prompt design or category confusion. Haiku receives the same multi-context prompts as the open-source models, yet assigns ENTITY to only 124 of 174,973 tokens (0.07%), compared to 47,276 from the majority vote, effectively defaulting all uncertain cases to NONE as instructed by the prompt. Across all tokens, it assigns NONE to 89.4%. This pattern places Haiku at the extreme end of a conservatism spectrum already visible among the open-source models, where Mistral (59.9% NONE) is markedly more conservative than Qwen (37.9%) and LLaMA (37.7%). Table 3 reports the full label distribution across all four models and the majority vote. The most striking pattern is Haiku's near-total rejection of the ENTITY class: of 47,276 majority-vote entities, none are confirmed and 97.3% are relabeled NONE. The verification stage thus acts as a strict precision filter, reducing the candidate set from 10,499 to 1,021.

6.4. Gold Standard Evaluation

The first author manually annotated all 1,021 pipeline output candidates using the same four-

Label	Qwen 72B	Mistral 123B	LLaMA 70B	Maj. vote	Haiku
NEOLOGISM	13,661 (7.8%)	11,493 (6.6%)	21,353 (12.2%)	10,499 (6.0%)	897 (0.5%)
ENTITY	56,144 (32.1%)	32,311 (18.5%)	55,088 (31.5%)	47,276 (27.0%)	124 (0.1%)
FOREIGN	38,851 (22.2%)	26,441 (15.1%)	32,625 (18.6%)	33,159 (19.0%)	17,506 (10.0%)
NONE	66,317 (37.9%)	104,728 (59.9%)	65,907 (37.7%)	84,039 (48.0%)	156,446 (89.4%)

Table 3: Label distribution per model and majority vote across all 174,973 tokens (count and % of total). LLaMA's NONE count includes 954 unparseable responses.

Gold label	Count	%
Lexical innovation	599	58.7
of which NEOLOGISM	465	45.5
of which ENTITY	134	13.1
Non-neologism	422	41.3
of which FOREIGN	61	6.0
of which NONE	361	35.4
Total	**1,021**	**100**

Table 4: Gold annotation of the 1,021 pipeline output candidates.

Process	Examples
Analogical formations	
Extra-gramm.: surface analogy	*updoot, pawrents*
Extra-gramm.: analogy via schema	
Secreted c.f.: *-fluencer*	*finfluencer, fitfluencer*
Abbreviated c.f.: *trad-*	*tradwife, tradfem*
Non-analogical formations	
Grammatical: prefixation	*deplatform, exvegan*
Grammatical: suffixation	*wokeism, trumpism*
Grammatical: compounding	*deepfake, longcovid*
Marginal: neoclassical c.f.	*abrosexual, acephobia*
Extra-gramm.: blending	*barbenheimer, maskne*
Extra-gramm.: expressive morph.	*thiccest, consoomer*

Table 5: Word-formation processes among gold neologisms.

class taxonomy, following the annotation criteria derived from the theoretical framework in §3: tokens were classified as NEOLOGISM if they resulted from a word-formation process (grammatical or extra-grammatical in the sense of Mattiello 2013) and were first attested after 2015, as ENTITY if they denoted a proper noun first attested after 2015, as FOREIGN if they belonged to another language, and as NONE otherwise. Table 4 cross-tabulates pipeline output against gold labels.

Of the 1,021 candidates, 599 (58.7%) are genuine lexical innovations: 465 neologisms and 134 named entities. The remaining 422 consist of 361 false positives (NONE) and 61 foreign-language tokens that escaped both rule-based and LLM-based detection.

7. Discussion

The pipeline is best understood as a high-recall candidate generator rather than a precision classifier. Its primary contribution is the 122,031:1 compression ratio, which reduces a task that no human annotator could feasibly undertake (reviewing 124.6 million tokens) to one that a single annotator can complete (reviewing 1,021 candidates). We do not report corpus-level recall, as the gold standard covers only the pipeline output; the number of neologisms in the 124.6 million tokens that the pipeline may have missed is unknown. To give a rough idea, however, an estimate based on an external reference list is provided in §7.3. This framing aligns with how comparable systems are evaluated: Tomaszewska et al. (2025) report precision at each stage and note that recall is not computable; Grieve et al. (2018) and Brasolin et al. (2023) similarly report counts of emerging words found. As in those systems, a final manual verification step is an integral part of the design, not a limitation.

7.1. Word-Formation Patterns

The 599 gold lexical innovations exhibit a range of word-formation processes that connect directly to the theoretical frameworks in §3. Table 5 organises the most productive patterns along two axes: whether the formation is analogical or non-analogical, and whether the process is grammatical, marginal, or extra-grammatical in the sense of Mattiello (2013).

Among non-analogical formations, standard grammatical processes account for a substantial share of the data. Prefixation with productive English prefixes yields forms such as *deplatform*, *detrash*, *exvangelical*, and *exvegan*. Suffixation with *-ism* generates *wokeism*, *trumpism*, *defaultism*, and *longtermism*. Compounding produces *deepfake*, *longcovid*, and *vibecheck*. At the margins of grammatical morphology, neoclassical combining forms of Latin or Greek origin appear in novel bases: *-sexual* (*abrosexual*, *dreamsexual*) and *-phobia/-phobic* (*acephobia*, *enbyphobic*).

Non-analogical extra-grammatical processes include blending, where two source words are fused without following a prior model (*barbenheimer*, *maskne*, *trumpanzee*), and expressive morphology, where deliberate phonological distortion of existing words produces new forms (*thiccest*, *consoomer*, *chonkster*).

The analogical formations are exclusively extra-grammatical. Surface analogy, where a single word serves as model, accounts for cases such as *updoot* (after *upvote*) and *pawrents* (after *parents*). More productive are formations arising through analogy via schema, where a recurrent fragment extracted from an initial blend becomes a combining form used across a series. Several such combining forms have undergone semantic generalisation, or secretion in the terminology of Mattiello (2013): *-fluencer* (from *influencer*; *finfluencer*, *fitfluencer*, *scamfluencers*), *-cel* (from *incel*; *femcel*, *mentalcels*), *-core* (from *hardcore*; *goblincore*, *traumacore*), *-nomics* (from *economics*; *bidenomics*, *tokenomics*), *-pilled* (from *redpilled*; *blackpilled*, *blackpillers*), and *-maxxing* (from *maxing*; *looksmaxxing*, *gymmaxxing*). Others function as abbreviated combining forms without semantic reinterpretation: *trad-* (from *traditional*; *tradwife*, *tradfem*) and *-flation* (from *inflation*; *greedflation*, *pissflation*). Whether the secreted forms have fully detached from their source words or remain at an intermediate stage between splinter and combining form is a diachronic question that the present data cannot resolve.

7.2. Error Analysis

The 422 non-neologistic tokens in the output fall into distinct categories, each pointing to a specific pipeline limitation.

False positives (361 tokens). The largest error category comprises concatenations (two or more words typed without a space, a common Reddit orthographic artifact) such as *datingapp*, *sidehustle*, and *telegramchannel* (approximately 50 tokens). These were correctly identified and removed by the word segmentation step, but subsequently reintegrated by the frequency threshold mechanism (§4.5), which restores all tokens with $\geq$100 occurrences regardless of the reason for their exclusion. At that point, the LLMs should have classified them as NONE, but failed to recognise them as mere orthographic variants of existing word sequences. A second cluster (approximately 50 tokens) comprises tokens from gaming and technical domains marked NONE for heterogeneous reasons. Some are names of programming functions or UI components (*floatlayout*, *floatmenu*, *floattensor*): accepting these would entail treating entire programming language vocabularies as natural language neologisms. Others are spaceless concatenations of pre-existing proper names (*bioniccommando*, a 1987 Capcom title; *biorepeel*, a cosmetic brand). Still others are fragments of game-internal proper names where the pipeline captured only part of a multi-token entity, or tokens that predate 2015 but were too domain-specific for the reference vocabulary. Approximately 30 tokens are misspellings of neologisms themselves (*neurodivegent*, *dollification*, *nuerotypicals*): the typo filter checks edit distance against *dictionary* words only and cannot detect that *neurodivegent* is a misspelling of *neurodivergent*, which is itself absent from the reference vocabulary. Finally, approximately 25 tokens are pre-2015 words absent from the 16.3 million-word reference vocabulary (*latinx*, *onfleek*, *biliteracy*): the vocabulary is comprehensive but not exhaustive for informal and slang registers.

Foreign language leakage (61 tokens). Two patterns dominate. Taglish (Tagalog–English) code-switching accounts for 29 of 61 tokens (47.5%): Tagalog prefixes (*na-*, *naka-*, *sina-*) affixed to English roots (*naghost*, *nakablock*, *sinasuggest*) superficially resemble English words with unfamiliar morphology, evading both Lingua and LLM classification. The remaining cases are English loanwords with Romance or Germanic inflection (*influenciador*, *influenceuse*, *brunchen*, *stressar*), originating from non-English posts where the mixed morphology falls below the language detector's confidence threshold.

The neologism–entity boundary. The 134 named entities in the gold standard highlight an inherently fuzzy boundary. Tokens such as *superstonk* (the subreddit name that became synonymous with the GameStop movement) and *barbenheimer* (*Barbie* + *Oppenheimer*) denote specific referents while also exhibiting productive word-formation processes (compounding, blending), making the neologism–entity distinction a matter of annotation judgment rather than a clear-cut category. Game-specific terms from *Splatoon* (Nintendo, 2015), numbering 15 tokens, *Among Us*, and various crypto projects account for the bulk of entities, and many are concentrated in one or two subreddits, a pattern that a cross-subreddit dispersion threshold could help address (see the Limitations section).

7.3. Pipeline False Negatives

To estimate recall, we compile a reference list of 103 single-token neologisms documented after 2015 in major dictionaries and lexicographic sources: Merriam-Webster additions (2016–2025),

Oxford and Collins Words of the Year, the American Dialect Society Word of the Year, the British Council "90 Words" list, Cambridge Dictionary, the OED, and Wiktionary, as well as community-maintained documentation sources such as Know Your Meme (full list in Appendix B). Of the 103 reference items, 20 are correctly detected; 48 were already attested on Reddit before 2015 and are correctly excluded; and 2 are excluded from evaluation as inflected forms of detected base forms. Loanwords (e.g. *mukbang* from Korean) are excluded, since borrowing falls outside the scope of the word-formation frameworks adopted in §3 (see the Limitations section). Inflected forms of base forms already detected by the pipeline (e.g. *deepfakes* alongside *deepfake*) are likewise excluded from the false negative count, as the pipeline's purpose is to identify novel lexical items rather than to capture every inflectional variant. The 33 genuine false negatives fall into the following categories: vocabulary homograph conflicts (17 tokens), where sparse pre-2015 occurrences in unrelated senses block the neologism (e.g. *rizz* as a character name, *simp* as a gaming clan); external vocabulary matches (9 tokens), where WordNet, Wiktionary, or Wikipedia contain the word under a different meaning (*copium*, *doggo*, *stonks*), a problem closely related to the semantic shift limitation discussed below; concatenation detection (3 tokens), where the segmentation module splits the token into known substrings (e.g. *cottagecore*); and the remaining 4 tokens are lost to tokenisation, typo correction, or LLM misclassification. Note that *doggo*, used in §3.2 as a canonical example of extra-grammatical morphology, is missed precisely because Wiktionary lists it under its pre-existing adverbial sense. Recall over the 53 genuinely post-2015 items is 20/53 (37.7%). Conversely, *vtuber* (virtual YouTuber), initially flagged as a typo of *tuber* by SymSpell, was correctly reintegrated by the frequency threshold mechanism owing to its 41,024 occurrences, illustrating that the reintegration stage (§4.5) functions as an effective safety net for high-frequency neologisms. The main axis for improvement is therefore conservative refinement of the vocabulary filtering stage, where type-level matching without sense disambiguation remains the primary source of false negatives.

8. Conclusion

We presented a scalable pipeline for automatic neologism detection that combines rule-based filtering with multi-model LLM classification, grounded in grammatical and extra-grammatical word-formation theory. Applied to 527 million Reddit posts, the pipeline achieves a 122,031:1 compression ratio, yielding 1,021 candidates of which 599 (58.7%) are genuine lexical innovations. Manual analysis of the output reveals a range of productive word-formation processes, from standard prefixation and compounding to analogical patterns such as secreted combining forms, confirming that the pipeline captures theoretically meaningful variation. The error analysis indicates that future improvements should target the earliest filtering stages rather than downstream classification. The pipeline code, vocabulary compilation scripts, and the annotated candidate list are available at https://github.com/DiegoRossini/neologism-pipeline.

Ethics Statement

The pipeline operates on unmoderated social media data and the resulting candidate list inevitably contains tokens related to offensive language, sexual content, hate speech, and extremist ideologies. Their inclusion reflects the lexical productivity of these domains and does not imply endorsement. All data was processed at the token level; no individual users were identified or tracked.

Limitations

The pipeline detects only single-token neologisms. Multi-word expressions such as *rage bait* or *touch grass* are invisible to the current architecture because the tokenizer treats each word independently, and no downstream stage attempts to reassemble multi-token units. Multi-word expressions are a major vector for lexical innovation in informal registers, and their absence from the output means the pipeline systematically underrepresents phrasal coinages.

Neologisms containing numerals or non-alphabetic characters are likewise excluded: the pattern cleaning stage (§4.3) discards all non-purely-alphabetic tokens. This filters out an increasingly productive category of lexical innovation known as algospeak (Steen et al., 2023; Aleksic, 2025), where users deliberately substitute letters with numbers or symbols to evade algorithmic content moderation (e.g., *$3X* for *sex*, *$trippers* for *strippers*), as well as named entities whose orthography includes digits, such as *4chan*.

The pipeline also cannot detect semantic shifts, where an existing word acquires a new meaning without any change in form. A prominent example is *Karen*, a conventional given name that underwent pejoration on Reddit and Black Twitter during the mid-2010s to denote an entitled, privileged white woman who weaponizes her social position. Because *Karen* is already present in the reference vocabulary as a proper noun, the pipeline excludes it at the vocabulary filtering stage, and no subsequent stage is equipped to detect that its usage distribution has changed.

The current instantiation targets English only. While the architecture is modular and transferable, all filtering resources, phonotactic rules, and frequency dictionaries are English-specific, and the LLM prompts are written in English. Adapting the pipeline to other languages would require substituting these components and re-evaluating the filtering thresholds. For morphologically rich languages with extensive inflectional paradigms, surface-form vocabulary matching may require either substantially larger observed-form vocabularies or an additional lemmatization step, though lemmatizing neologisms is itself problematic, since a lemmatizer trained on existing vocabulary may not reliably reduce novel forms to their base.

The foreign language detection stage (§4.6) cannot distinguish non-English corpus noise from genuine loanwords entering English. Lexical borrowing is not a word-formation process in either framework adopted in §3 — neither Štekauer (2002) nor Mattiello (2013) include it in their taxonomies, consistent with the position that borrowing and word formation are fundamentally distinct, though interacting, domains (ten Hacken and Panocová, 2020). This means that nativised loanwords such as *mukbang* or *hygge*, which are established terms in English, fall outside the pipeline's scope and are excluded either by the language detector or by the reference vocabulary.

As discussed in §7.3, the most consequential false negatives originate at the tokenization and vocabulary filtering stages, where tokens attested in pre-2015 data as probable typos (*stonk*, *monke*) or present in encyclopaedic sources (*copium*) are silently treated as known vocabulary and never enter the candidate pool. Two targeted improvements would address recurrent false positive patterns identified in §7.2. A cross-subreddit dispersion threshold would complement the raw frequency threshold by requiring candidates to appear across a minimum number of distinct subreddits, filtering concatenations such as *datingapp* that accumulate high frequencies within a single community through repeated orthographic error rather than deliberate coinage. A post-classification deduplication step comparing candidates by edit distance would catch misspellings of neologisms already captured by the pipeline (e.g., *neurodivegent* alongside *neurodivergent*).

All filtering thresholds (e.g. token length, edit distance, frequency, language detection confidence) were set based on preliminary experimentation rather than systematically optimised on a development set, as the computational cost of a full pipeline run (50–65 hours) makes exhaustive parameter search impractical. A systematic sensitivity analysis is left for future work.

Data and Code Availability

The pipeline code, vocabulary compilation scripts, and the annotated candidate list are available at https://github.com/DiegoRossini/neologism-pipeline.

Acknowledgements

This research was funded by the NCCR Evolving Language, Swiss National Science Foundation Agreement No. 51NF40_225146. We thank the anonymous reviewers for their helpful feedback, Ryan of NoSlang.com for generously sharing the abbreviation list, and Quirin Würschinger for personal communication regarding the NeoCrawler.

9. Bibliographical References

Adam Aleksic. 2025. *Algospeak: How Social Media Is Transforming the Future of Language*. Knopf, New York.

Sabine Arndt-Lappe. 2015. Word-formation and analogy. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation: An International Handbook of the Languages of Europe*, volume 2, pages 822–841. De Gruyter Mouton, Berlin.

Geert Booij. 2010. *Construction Morphology*. Oxford University Press, Oxford.

Paolo Brasolin, Greta H. Franzini, and Stefania Spina. 2023. "Ti blocco perché sei un trollazzo": Lexical Innovation in Contemporary Italian in a Large Twitter Corpus. *Journal of Italian Linguistics*, 35(2):123–145.

Maria Teresa Cabré and Lluís de Yzaguirre. 1995. Stratégie pour la détection semi-automatique des néologismes de presse. *TTR: Traduction, Terminologie, Rédaction*, 8(2):89–100.

Emmanuel Cartier. 2017. Neoveille, a web platform for neologism tracking. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 95–98, Valencia, Spain. Association for Computational Linguistics.

Wolfgang U. Dressler. 2000. Extragrammatical vs. marginal morphology. In Ursula Doleschal and Anna M. Thornton, editors, *Extragrammatical and Marginal Morphology*, number 12 in LINCOM Studies in Theoretical Linguistics, pages 1–10. Lincom Europa, München.

Ingrid Falk, Delphine Bernhard, and Christophe Gérard. 2014. From non word to new word: Automatically identifying neologisms in French newspapers. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 4337–4344, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jack Grieve, Andrea Nini, and Diansheng Guo. 2018. Mapping Lexical Innovation on American Social Media. *Journal of English Linguistics*, 46(4):293–319.

Louis Guilbert. 1975. *La créativité lexicale*. Langue et Langage. Larousse, Paris.

Peter Hohenhaus. 1998. Non-lexicalizability as a characteristic feature of nonce word-formation in English and German. *Lexicology*, 4(2):237–280.

Daphné Kerremans, Jelena Prokić, Quirin Würschinger, and Hans-Jörg Schmid. 2018. Using data-mining to identify and study patterns in lexical innovation on the web: The *NeoCrawler*. *Pragmatics & Cognition*, 25(1):174–200.

Daphné Kerremans, Susanne Stegmayr, and Hans-Jörg Schmid. 2012. The NeoCrawler: Identifying and Retrieving Neologisms from the Internet and Monitoring Ongoing Change. In *Current Methods in Historical Semantics*, pages 59–96.

Annette Klosa-Kückelhaus and Harald Lüngen. 2018. New German words: Detection, description, and dictionary entry. In *Lexicography in the Digital Age*, pages 559–569. Euralex.

Lívia Körtvélyessy, Pavol Štekauer, and Pavol Kačmár. 2021. On the role of creativity in the formation of new complex words. *Linguistics*, 59(4):1017–1055.

Lívia Körtvélyessy, Pavol Štekauer, and Pavol Kačmár. 2022. *Creativity in Word Formation and Word Interpretation: Creative Potential and Creative Performance*, 1 edition. Cambridge University Press.

Taylor Mahler. 2020. Lexical Emergence on Reddit. *Lexis – Journal in English Lexicology*, 16.

Elisa Mattiello. 2013. *Extra-Grammatical Morphology in English: Abbreviations, Blends, Reduplicatives, and Related Phenomena*. Number 82 in Topics in English Linguistics. De Gruyter Mouton, Berlin.

Elisa Mattiello. 2017. *Analogy in Word-formation: A Study of English Neologisms and Occasionalisms*. Number 309 in Trends in Linguistics. Studies and Monographs. De Gruyter Mouton, Berlin.

Ingo Plag. 1999. *Morphological Productivity: Structural Constraints in English Derivation*. Mouton de Gruyter, Berlin.

Antoinette Renouf. 1993. A word in time: First findings from dynamic corpus investigation. In Jan Aarts, Pieter de Haan, and Nelleke Oostdijk, editors, *English Language Corpora: Design, Analysis and Exploitation*, pages 279–288. Rodopi, Amsterdam.

Alain Rey. 1976. Néologisme: un pseudo-concept? *Cahiers de Lexicologie*, 28(1):3–17.

Stefania Spina, Paolo Brasolin, and Greta H. Franzini. 2024. Detecting emerging vocabulary in a large corpus of Italian tweets. *Research in Corpus Linguistics*, 13(1):139–170.

Ella Steen, Kathryn Yurechko, and Daniel Klug. 2023. You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on TikTok. *Social Media + Society*, 9(3).

Pavol Štekauer. 2001. Fundamental principles of an onomasiological theory of English word-formation. *Onomasiology Online*, 2:1–42.

Pius ten Hacken and Renáta Panocová, editors. 2020. *The Interaction of Borrowing and Word Formation*. Edinburgh University Press, Edinburgh.

Aleksandra Tomaszewska, Dariusz Czerski, Bartosz Żuk, and Maciej Ogrodniczuk. 2025. NeoN: A Tool for Automated Detection, Linguistic and LLM-Driven Analysis of Neologisms in Polish. ArXiv:2505.15426 [cs].

Quirin Würschinger. 2021. Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter. *Frontiers in Artificial Intelligence*, 4:648583.

Arnold M. Zwicky and Geoffrey K. Pullum. 1987. Plain morphology and expressive morphology. In *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, pages 330–340.

Pavol Štekauer. 1998. *An Onomasiological Theory of English Word-Formation*, volume 46 of *Studies in Functional and Structural Linguistics*. John Benjamins Publishing Company, Amsterdam.

Pavol Štekauer. 2002. On the Theory of Neologisms and Nonce-formations. *Australian Journal of Linguistics*, 22(1):97–112.

Pavol Štekauer. 2005. Onomasiological Approach to Word-Formation. In Marcel Den Dikken, Liliane Haegeman, Joan Maling, Guglielmo Cinque, Carol Georgopoulos, Jane Grimshaw, Michael

Kenstowicz, Hilda Koopman, Howard Lasnik, Alec Marantz, John J. McCarthy, Ian Roberts, Pavol Štekauer, and Rochelle Lieber, editors, *Handbook of Word-Formation*, volume 64, pages 207–232. Springer Netherlands, Dordrecht. Series Title: Studies in Natural Language and Linguistic Theory.

10. Language Resource References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset.

Wolf Garbe. 2012. SymSpell: Symmetric delete spelling correction algorithm.

Princeton University. 2011. WordNet 3.1.

Peter M. Stahl. 2022. Lingua: The most accurate natural language detection library for python.

Urban Dictionary. 2025. Urban Dictionary entry database. Scraped and filtered to entries predating 2015.

Watchful1. 2025. Subreddit comments/submissions 2005-06 to 2024-12. Per-subreddit split of the Pushshift Reddit dumps. Available via Academic Torrents.

Wikimedia Foundation. 2015a. Wikipedia: English article titles dump. Dump dated 2015-01-01.

Wikimedia Foundation. 2015b. Wiktionary: English edition dump. Dump dated 2015-01-01.

A. Prompt Templates

Both prompts are used identically across all four models (Qwen 72B, LLaMA 70B, Mistral Large 123B, and Claude Haiku).

Multi-token prompt (primary pass, 10 tokens per call).

```
TASK: Classify each token
into ONE category.

ENTITY - Pure proper nouns only
  (real/fictional): people, characters,
  companies, brands, products, games,
  movies, places, apps
  Examples: elon, pikachu, google,
  iphone, fortnite, reddit, tokyo

NEOLOGISM - New English words, slang,
  OR words derived from proper nouns
  Examples: doomscrolling, ghosting,
  rizz, bussin, adulting, covidiot,
  youtuber, redditor, trumpian,
  instagrammable, uberize, googlable

FOREIGN - Non-English words
  Examples: além, anspielung,
  yapmyorum, además

NONE - Usernames, typos, programming
  terms, unclear words

CRITICAL RULES:
1. Derived forms are NEOLOGISM
   (youtuber -> NEOLOGISM,
    youtube -> ENTITY)
2. When uncertain, classify as NONE
3. Use the context and subreddit
   to understand usage

TOKENS:
TOKEN: <token_1>
  context_1 (r/<subreddit>): "<text>"
  context_2 (r/<subreddit>): "<text>"
  context_3 (r/<subreddit>): "<text>"
TOKEN: <token_2>
  context_1 (r/<subreddit>): "<text>"
  ...

OUTPUT:
One classification per line as
TOKEN:LABEL (ENTITY, NEOLOGISM,
FOREIGN, or NONE).
No explanations.
```

Single-token prompt (retry pass for failed tokens).

```
Classify this token into ONE
category: ENTITY, NEOLOGISM,
FOREIGN, or NONE.

ENTITY - Pure proper nouns only
  (real/fictional): people,
  characters, companies, brands,
  products, games, movies, places,
  apps
NEOLOGISM - New English words, slang,
  OR words derived from proper nouns
  (youtuber, trumpian,
   instagrammable)
FOREIGN - Non-English words
NONE - Usernames, typos, programming
  terms, unclear words

TOKEN: <token>
  context_1 (r/<subreddit>): "<text>"
  context_2 (r/<subreddit>): "<text>"
  context_3 (r/<subreddit>): "<text>"

Answer with ONLY the label:
<token>:LABEL
```

B. Recall Reference List

Table 6 and Table 7 list the 103 single-token neologisms used for the recall evaluation in §7.3.

Status labels. TP = detected by the pipeline (true positive); FN = genuine false negative (post-2015, missed by pipeline); pre-15 = correctly excluded (attested on Reddit before 2015); excl. = excluded from evaluation (inflected form of a detected base form).

Source abbreviations. MW = Merriam-Webster; KYM = Know Your Meme; BC90 = British Council 90 Words; ADS = American Dialect Society; Collins = Collins Dictionary; Cambridge = Cambridge Dictionary; Oxford WOTY = Oxford Word of the Year; UrbanDict = Urban Dictionary; Dictionary.com = Dictionary.com; Aesth. Wiki = Aesthetics Wiki.[2]

[2]Source base URLs: Merriam-Webster: https://www.merriam-webster.com; Know Your Meme: https://knowyourmeme.com; British Council 90 Words: https://www.britishcouncil.org; American Dialect Society: https://www.americandialect.org; Collins Dictionary: https://www.collinsdictionary.com; Cambridge Dictionary: https://dictionary.cambridge.org; Oxford Word of the Year: https://languages.oup.com/word-of-the-year; Urban Dictionary: https://www.urbandictionary.com; Dictionary.com: https://www.dictionary.com; Wiktionary: https://en.wiktionary.org; Aesthetics Wiki: https://aesthetics.fandom.com. The full reference list with per-word verification URLs is available in the project repository.

Word	Year	Source	Status
doomscroll	2020	MW 2023	TP
doomscrolling	2020	MW 2023	TP
deepfake	2017	MW 2023; BC90	TP
deepfakes	2017	MW 2023	excl.
finsta	2017	MW 2023	FN
edgelord	2016	MW 2023; BC90	pre-15
copypasta	2016	MW 2023	pre-15
clickbait	2015	MW 2018	pre-15
subtweet	2015	MW 2018	pre-15
doxing	2015	MW 2017	pre-15
doxxing	2015	MW 2023	pre-15
ghosting	2017	MW 2017	pre-15
catfishing	2015	MW 2023	pre-15
copium	2020	Collins; MW	FN
hopium	2020	Collins	pre-15
shitposting	2017	Wiktionary	pre-15
shitpost	2017	Wiktionary	pre-15
rizz	2023	MW 2023; Oxford WOTY 2023; BC90	FN
simp	2019	MW 2023	FN
simping	2019	MW 2025	FN
stan	2017	MW 2019	pre-15
stanning	2017	MW 2019	pre-15
sealioning	2017	Collins	pre-15
doggo	2017	MW 2023	FN
birb	2017	KYM	FN
chonk	2018	KYM	FN
chonky	2018	KYM	FN
poggers	2017	KYM	FN
stonks	2021	KYM	FN
thicc	2017	KYM	FN
updoot	2016	KYM	TP
yeet	2018	MW 2023	pre-15
yeeted	2018	MW 2023	FN
sussy	2021	KYM	FN
bussin	2021	MW 2023	pre-15
skibidi	2023	Cambridge 2025	FN
delulu	2023	Cambridge 2025	FN
uwu	2017	KYM	pre-15
smol	2016	KYM	FN
blorbo	2022	KYM	TP
enshittification	2023	ADS WOTY 2023	TP
enshittify	2023	Wiktionary	FN
touchgrass	2021	MW 2024	FN
blockchain	2016	MW 2018	pre-15
cryptocurrency	2017	MW 2018	pre-15
bitcoin	2016	MW 2016	pre-15
chatbot	2017	MW 2018	pre-15
ransomware	2017	MW 2018	pre-15
deepfaked	2019	Wiktionary	FN
vtuber	2020	Wiktionary	TP
hodl	2017	Wiktionary	pre-15
defi	2020	Wiktionary	pre-15

Table 6: Recall reference list (1/2). "Year" indicates when the source documented the word, not the year of coinage.

Word	Year	Source	Status
altcoin	2017	Wiktionary	pre-15
memecoin	2021	Wiktionary	pre-15
stablecoin	2020	Wiktionary	pre-15
rugpull	2021	Wiktionary	FN
rugpulled	2021	Wiktionary	FN
wokeism	2019	Wiktionary	TP
wokeness	2019	Wiktionary	FN
trumpism	2016	Wiktionary	TP
deplatform	2018	Wiktionary	TP
deplatformed	2018	Wiktionary	excl.
deplatforming	2018	Wiktionary	TP
mansplaining	2015	MW 2018	pre-15
manspreading	2015	MW 2016	pre-15
whataboutism	2017	MW 2019	pre-15
incel	2018	Collins WOTY 2018; BC90	pre-15
incels	2018	Collins WOTY 2018	pre-15
blackpill	2018	Wiktionary	FN
blackpilled	2018	Wiktionary	TP
redpilled	2016	Wiktionary	pre-15
breadcrumbing	2018	Wiktionary	TP
situationship	2022	Oxford WOTY 2023; BC90	pre-15
allyship	2018	MW 2019	pre-15
covidiot	2020	Collins	TP
quarantini	2020	Wiktionary	FN
longcovid	2020	Wiktionary	TP
superspreader	2020	MW 2020	FN
doomscroller	2020	MW 2023	TP
covfefe	2017	Wiktionary	TP
infodemic	2020	Wiktionary	FN
deadname	2018	MW 2023	FN
deadnaming	2018	MW 2023	FN
genderfluid	2016	MW 2018	pre-15
demisexual	2018	Wiktionary	pre-15
neurodivergent	2020	MW 2023	pre-15
adulting	2016	MW 2017; BC90	pre-15
cottagecore	2020	Dictionary.com	FN
goblincore	2020	Dictionary.com	TP
darkcore	2020	Aesth. Wiki	pre-15
tradwife	2020	Cambridge 2025	TP
sponcon	2018	Wiktionary	FN
finfluencer	2020	Wiktionary	TP
hygge	2016	Oxford WOTY 2016	pre-15
glamping	2016	MW 2016	pre-15
athleisure	2016	MW 2016	pre-15
shadowban	2022	MW 2024	pre-15
jawnz	2023	UrbanDict	pre-15
longhauler	2020	Wiktionary	FN
rawdogging	2024	Wiktionary	pre-15
brainrot	2024	Oxford WOTY 2024	FN
airdrop	2017	Wiktionary	pre-15
gaslighting	2016	MW WOTY 2022	pre-15

Table 7: Recall reference list (2/2). "Year" indicates when the source documented the word, not the year of coinage.