NEO-BENCH: Evaluating Robustness of Large Language Models with Neologisms

Anonymous ACL submission

Abstract

The performance of Large Language Models (LLMs) degrades from the temporal drift between data used for model training and newer text seen during inference. One understudied avenue of language change causing data drift is the emergence of neologisms – new word forms – over time. We create a diverse resource of recent English neologisms by using several popular collection methods. We analyze temporal drift using neologisms by comparing sentences containing new words with near-identical sentences that replace neologisms with existing substitute words. Model performance is nearly halved in machine translation when a single neologism is introduced in a sentence. Motivated by these results, we construct a benchmark to evaluate LLMs' ability to generalize to neologisms with various natural language understanding tasks and model perplexity. Models with later knowledge cutoff dates yield lower perplexities and perform better in downstream tasks. LLMs are also affected differently based on the linguistic origins of words, indicating that neologisms are complex for static LLMs to address. We will release our benchmark and code for reproducing our experiments.

1 Introduction

Neologisms – recent word forms representing a new meaning, sense, or connotation (Cartier, 2017) – consistently surface as language changes. Neologisms emerge to describe the ever-changing state of the world, such as new terms created during the COVID-19 pandemic. While humans easily adapt to language change, large language models (LLMs) struggle with the misalignment of training data and new test data distributions (Luu et al., 2022).

Prior work on temporal language change (Lazaridou et al., 2021; Onoe et al., 2022; Luu et al., 2022) observed model degradation when finetuning on older text and evaluating on newer data and named entities (Rijhwani and Preotiuc-Pietro, 2020; Agarwal and Nenkova, 2022; Liu and Ritter, 2023). However, as far as we are aware there has not been prior work that analyzes the robustness of LLMs on handling neologisms. We show that adding a neologism to text decreases machine translation quality by an average of 44% in a human evaluation (§2), even for popular words emerging before 2020.

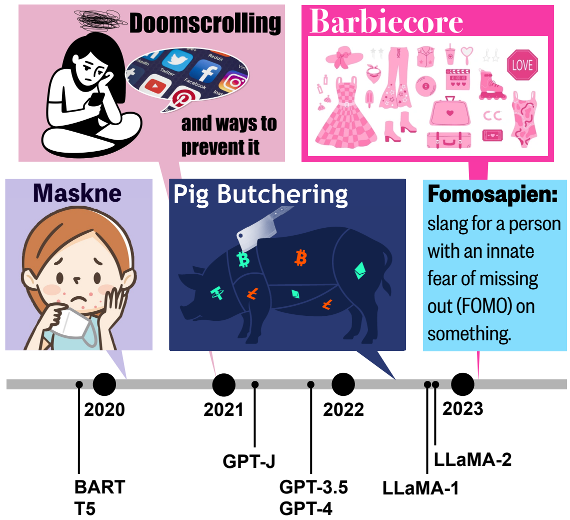

Figure 1: NEO-BENCH collects neologisms from 2020-2023 for LLM evaluation. "Pig Butchering" originated as a Mandarin expression (杀猪盘).

In this paper, we present NEO-BENCH, a new benchmark designed to test the ability of LLMs to understand and process neologisms. We combine multiple methods and online text corpora to collect a diverse set of 2,505 neologisms based on the linguistic taxonomy devised by Pinter et al. (2020): (i) **lexical neologisms** – words representing new concepts, e.g., *"long covid"*; (ii) **morphological neologisms** – blends of existing subwords, e.g., *"doomscrolling"*; and (iii) **semantic neologisms** – existing words that convey a new meaning or sense, e.g., *"ice"* (a term that refers to petrol- or diesel-powered cars taking electric car charging spots). We estimate word prevalence over time with Google Trends to obtain trending neologisms. We also create 4 benchmark tasks to evaluate the

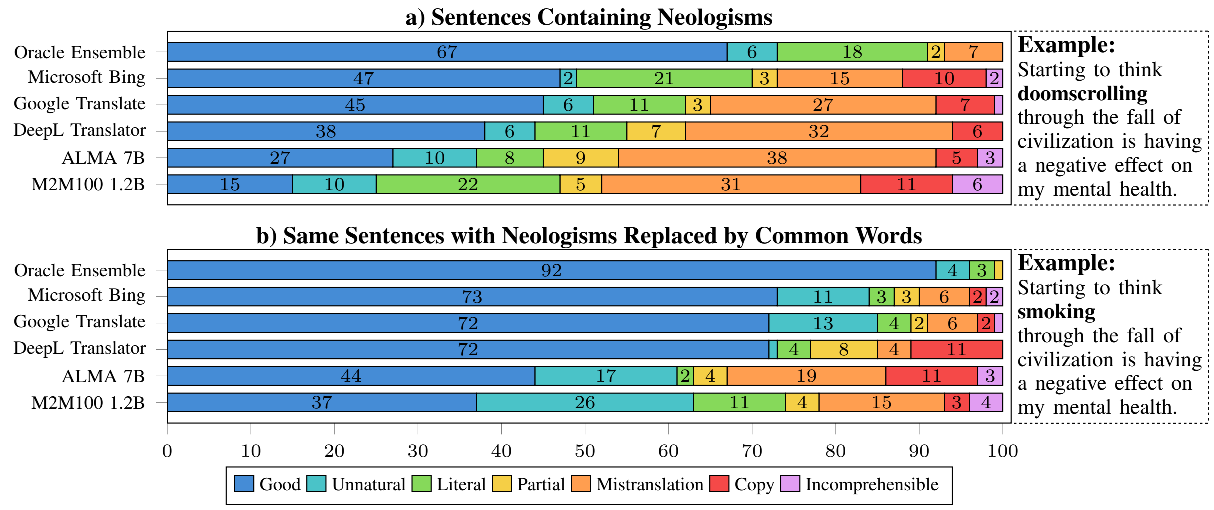

Figure 2: A single neologism can dramatically affect model output, as shown by human evaluation of Machine Translation models on sentences containing neologisms and the same sentences with neologisms replaced by carefully chosen words that also fit in the context. Oracle ensemble selects the best translation from all models.

impact of neologisms on LLMs with Perplexity, Cloze Question Answering, Definition Generation, and Machine Translation.

We show that lower neologism perplexities correlate with higher downstream task performance. Older LLMs – BART, T5, GPT-J, and Flan-T5 – perform much worse with an average of 32.20% and 12.27% accuracy in question answering and definition generation, respectively. We also find that automatic metrics do not accurately measure the quality of translated sentences containing neologisms, evidenced by Spearman's ρ rank correlation between COMET-Kiwi (a state-of-the-art metric) and human judgment, which is 0.508. This is lower than the average ρ of 0.629 for COMET-Kiwi across 5 language pairs reported in the WMT23 Quality Estimation task (Blain et al., 2023). LLM performance in NEO-BENCH also differs based on a word's linguistic type, as lexical neologisms without derivations yield the highest perplexities and the most fragmented subword tokenization, while semantic neologisms that repurpose existing words result in literal definitions and translations.

NEO-BENCH evaluates a diverse set of LLM capabilities on handling neologisms in various tasks. Models must also understand compositionality for morphological neologisms, differentiate between word senses for semantic neologisms, and handle different contexts for lexical neologisms.

2 Motivation

We start by using machine translation as an example to illustrate the significant challenge neologisms pose on state-of-the-art NLP systems. We manually collect 100 neologism words with sentential context from social media, news articles, and new dictionary entries. The best commercial translation systems, e.g., Google Translate,[1] Microsoft Bing,[2] and DeepL Translator,[3] only managed to correctly translate about 38-47% of these 100 sentences that contain neologisms based on our manual inspection (Figure 2; from English to Chinese). In stark contrast, if we replace that one neologism word with a common word in these sentences, the percentage of correct translations rises substantially to 72-73%. We observe similar trends in open-source translation models, such as ALMA (Xu et al., 2023) and M2M100 (Fan et al., 2020).

One thing to note is that these replacement words are not exact synonyms, but words that have been carefully chosen to create a near-identical, semantically plausible sentence; because new words emerge in areas not occupied by existing words (Ryskina et al., 2020), true synonyms would often be verbose and incompatible with the sentence context. Because the original sentences containing neologisms were collected in the wild, one might assume they would be even more natural in comparison to their modified counterparts, but yet, **there is a large gap in translation quality between neologism and non-neologism words for all models**.

A closer look reveals that six typical types of errors are made in mistranslated model outputs, which include (ordered by severity):

- **Unnatural**: Imperfect translation of the sen-

[1] https://translate.google.com/
[2] https://www.bing.com/translator
[3] https://www.deepl.com/translator

tence due to grammatical errors;

- **Literal**: Inaccurate output that literally translates the neologism or remaining sequence;
- **Partial**: Part of the sentence is untranslated and left out of the output;
- **Mistranslation**: Incorrectly translated sentence portion leads to a poor understanding of the overall sentence meaning;
- **Copy**: Part of the output is not translated and copied from the English input;
- **Incomprehensible**: Incoherent output that fails to capture any original sentence meaning;

Table 10 in the Appendix shows translations for each error type. The most common errors are mistranslations and literal translations with an average of 28.6% and 14.6% respectively. Model output for non-neologism sentences is more likely to have minor errors and be labeled unnatural by annotators.

Another interesting observation is that newer neologisms indeed show lower rates of good translations and often higher rates of mistranslations, as one may expect. Figure 3 shows the percentage of good translations and mistranslations over time for varied models. Compared to non-neologism sentences, models still yield lower rates of correct translations for neologisms that emerged before 2020. Many neologisms use existing words to convey meanings, such that the poor performance of models is not wholly explained by the absence of these word forms in training data. We propose a novel benchmark (§3) to systematically study the impact of neologisms on LLMs (§4).

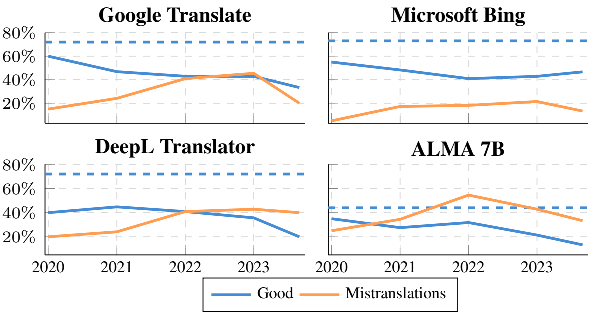

Figure 3: Percentage of good translations and mistranslations of neologism sentences over time. The dashed line represents the percentage of good translations achieved on non-neologism sentences.

Task	Dataset	Evaluation
Machine Translation	240 sentences containing neologisms	BLEU, COMET
Perplexity Ranking	422 Cloze passages with one-word answers	Word ranking
Cloze Questions	750 Cloze passages with multiple choice answers	Accuracy
Definition Generation	750 "What is [neologism]?" questions	Accuracy

Table 1: Summary of datasets in NEO-BENCH.

3 NEO-BENCH: A Neologism Benchmark

We create NEO-BENCH, a benchmark that consists of 2,505 neologisms (both words and phrases) that newly emerged around 2020–2023 and 4 intrinsic/extrinsic tasks (Table 1) to evaluate LLMs' abilities to generalize on neologisms.

3.1 Neologism Collection

A neologism is a term that represents a new meaning or sense (Cartier, 2017). Previous datasets (McCrae, 2019; Ryskina et al., 2020; Zhu and Jurgens, 2021) only collected specific word types, ignored neologisms conveying new meanings with existing words, and did not utilize word prevalence trends (more in Related Work §6). We design a more systematic collection process to quantify the effect of neologisms on a language's data distribution.

Filtering Reddit Data based on Google Trends (Method 1). New words commonly propagate in online communities (Zhu and Jurgens, 2021), thus, we count word frequencies in monthly Reddit data to find single-word neologism candidates. We set a frequency cutoff between 50 and 100 per month to obtain uncommon words and remove misspellings and named entities using SpaCy (Montani et al., 2023), resulting in 74,542 candidates. We further obtain word search frequencies from 2010 to 2023 on Google Trends[4] and automatically filter out 87.13% of neologism candidates based on these trend lines (see Figure 4 for examples) by a combination of curve fitting, argmax detection, and integrals over time. Appendix §B.1 provides more details about trend filtering. From the set of 9,590 remaining candidates, we find that 10.48% are prevalent neologisms by manual inspection. In total, we collected 1,005 neologism words from Reddit (310 lexical, 588 morphological, and 107 semantic neologisms).

Retrieving News Articles about Neologisms (Method 2). As Method 1 is only good at finding single-word neologisms, we turn to news articles that explain the meanings of neologisms to collect multi-word expressions. We first manually

[4] https://trends.google.com/trends/

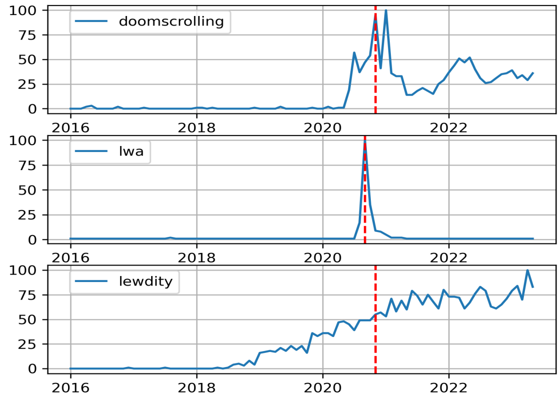

Figure 4: Example Google Trend lines measuring neologism prevalence. The dashed line estimates the date a neologism becomes popular while not yet conventional.

Neologism: ***doomscrolling***	
The silver lining of this website no longer functioning as an even vaguely reliable information source is that ___ has basically been completely undermined. It wouldn't even work now since everything is too geared to outrage clickbait and actual reporting has disappeared, so there is no point staying on the app.	
a) misinformation	*b) surfing*
c) doomscrolling	*d) lying*
e) gaming	***f) anti-productivity (distractor)***

Table 2: Example passage in NEO-BENCH for multiple-choice Cloze Question Answering with correct neologism answers and partially correct distractor answers.

get 100 neologisms from news articles, recording news headlines of neologisms. Then, based on the shared text patterns of headlines, we created 16 headline **templates** (e.g., *"___: What is it?"*) to retrieve Google News articles from 2019 to 2023. Using SpaCy, we identify 60,671 noun and verb phrases with a Part-of-Speech tagger and remove duplicates and named entities. We used the same aforementioned filtering method for these phrases using Google Trends. From the remaining 8,039 candidates, we manually extracted 1,100 neologisms (778 lexical, 222 morphological, and 100 semantic neologisms), of which 713 are multiwords.

Sampling Existing Neologism Datasets (Method 3). To supplement our dataset with additional neologisms, we also sample from two existing open-source resources that contain a lot of rare words, many of which have no Google Trends data available. The NYT First Said Twitter bot (@NYT_first_said) tweets out words when they are used for the first time in New York Times articles by using exclusion lists. We retrieve 1,100 of its tweets from 2020 to collect 200 derived neologisms (192 morphological and 8 semantic). We also sample 1,400 entries from another noisy, automatically constructed, dataset of 80,071 new slang dictionary entries (Zhu and Jurgens, 2021). We manually filter the sample and collect 200 derived neologisms (4 lexical, 194 morphological, and 2 semantic).

Overall. We collected 2,505 neologism words. While semantic neologisms are infrequent in all sources, Google Trends data enables the collection of them, as these words change in baseline prevalence when a new sense is being popularized. Only 5.04% of words from Reddit, 1.12% of phrases from news articles, and 3.09% of entries from previous datasets overlap with candidates from the other two sources — highlighting the importance of using multiple diverse data sources and methods for neologism collection. We also verified that 44.23% of these 2,505 words actually appear in the Urban Dictionary, a crowdsourced English-language online dictionary for slang words and phrases.

3.2 Benchmark Tasks

NEO-BENCH consists of 4 tasks – 3 downstream and 1 intrinsic metric – to evaluate models' knowledge of neologisms: (i) Machine Translation with human and automatic evaluation; (ii) Cloze Question Answering to evaluate models in context; (iii) Definition Generation to evaluate models in a context-free setting; and (iv) perplexity to compare single-word neologisms to commonly used words. We describe the setup and result tables/figures in this section, then discuss the key findings based on these results more in-depth in §4.

Machine Translation (Task 1). We sample from our collected neologisms (§3.1) and search for reference sentences containing these words on social media and Google. We construct 240 sentences, including the 100 used in §2. We work with in-house native speakers to create reference translations (English to Chinese) and evaluate system outputs in Table 4 with BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), and COMET-Kiwi (Rei et al., 2022), a reference-free metric. We also report the correlation of these metrics with human ratings for 5 MT models in §2 using Spearman's ρ in Table 4.

Cloze Question Answering (Task 2). We sample 750 neologisms to create text passages, where one sentence contains a neologism and the remaining passage serves as preceding or following context (see example in Table 2). We mask out the neologism and provide four incorrect answers plus one distractor answer, which is a common word or phrase that is feasible in context. We evaluate BART-large (Lewis et al., 2019), T5-Large (Raffel et al., 2020), Flan-T5-Large (Chung et al.,

a) Definition Generation Output Examples	
Stablecoin	**Reference Definition:** *A stablecoin is a type of cryptocurrency where the value of the digital asset is supposed to be pegged to a reference asset, which is either fiat money, exchange-traded commodities, or another cryptocurrency.*
	Model Output (Correct): *Stablecoins are cryptocurrencies designed to maintain a stable value, typically by pegging their value to a specific asset or basket of assets, such as the US dollar, gold, or a combination of assets.*
Angel Shot	**Reference Definition:** *An angel shot is a code to inform a bartender that a customer is not safe and needs assistance.*
	Model Output (Incorrect): *An angel shot is a cocktail made with whiskey and cream, served in a shot glass.*
b) Machine Translation Output Examples	
Longcovid	**Input:** Each reinfection increases the risk of **longcovid**, hospitalization, & death.
	Model Output (Correct): 每次再感染都会增加长新冠病毒、住院和死亡的风险。 *(Every reinfection increases the risk of* ***long COVID****, hospitalization, and death.)*
	Human Translation: 每一次新冠感染都会提高出现后遗症、住院治疗，甚至死亡的风险。 *(Each COVID-19 infection increases the risk of developing* ***sequelae****, hospitalization, and even death.)*
Doomscrolling	**Input:** Starting to think **doomscrolling** through the fall of civilization is having a negative effect on my mental health.
	Model Output (Incorrect): 开始认为在文明的衰落中滚动的厄运对我的心理健康产生了负面影响。 *(Start to think that the* ***doom rolling*** *in the decline of civilization is having a negative impact on my mental health.)*
	Human Translation: 开始觉得, 刷关于文明衰败的负能量新闻对我的心理健康产生了负面影响。 *(Starting to feel that* ***scrolling*** *through negative news about the decline of civilization is having a negative impact on my mental health.)*

Table 3: Example model definitions and translations for NEO-BENCH tasks. *"Doomscrolling"* is the act of spending an excessive amount of time reading negative news online. (English translations are shown for information only.)

2022), GPT-J 6B (Wang and Komatsuzaki, 2021), LLaMA-1 7B (Touvron et al., 2023a), Alpaca 7B (Taori et al., 2023), LLaMA-2, LLaMA-2-Chat (Touvron et al., 2023b), GPT 3.5 (Brown et al., 2020), and GPT-4 in multiple-choice Cloze Question Answering (QA). We experiment with 5-shot prompting and test three sizes of LLaMA-2 models. We show results in Figure 5 with the stratified and combined accuracies of selecting either the neologism or distractor answer.

Open-ended Definition Generation (Task 3). We evaluate the same models from Task 2 for their context-free knowledge of 750 neologisms with question prompts (i.e., *"What is doomscrolling?"*) to obtain neologism definitions. We construct human reference definitions and use GPT-4 to evaluate if model generations are semantically equivalent to the gold reference. We use 5-shot prompting and report results with accuracy in Figure 5. Table 3 shows example LLM-generated definitions.

Perplexity Rankings (Task 4). Using 422 Cloze passages that have both singular distractor and neologism answers, we use perplexity to evaluate GPT-J 6B, LLaMA-1 7B, Alpaca 7B, LLaMA-2 7B, and LLaMA-2 Chat 7B. For each passage, we use rank classification (Brown et al., 2020), where we fill in the mask with the neologism and measure the perplexity of the passage. We replace the mask with the distractor answer and the top 5000 singular words from Reddit by frequency and measure sequence perplexities. The mask-filling words are sorted by perplexity, and the average model rankings of neologisms and distractors are reported in Figure 7. Lower neologism rankings represent lower relative perplexities, indicating that a model is likely to complete the passage with a neologism.

Model (human rank)		**COMET**	**COMET-Kiwi**	**BLEU**
Bing Translator	(1)	0.825 (5)	0.788 (3)	0.452 (2)
Google Translate	(2)	**0.854** (1)	0.800 (2)	**0.507** (1)
DeepL Translator	(3)	0.842 (3)	**0.807** (1)	0.406 (4)
ALMA 7B	(4)	0.801 (8)	0.746 (8)	0.285 (7)
M2M100 1.2B	(5)	0.776 (9)	0.745 (9)	0.337 (6)
GPT-4		0.846 (2)	0.781 (5)	0.442 (3)
GPT-3.5		0.836 (4)	0.783 (4)	0.390 (5)
ALMA-R 13B		0.818 (6)	0.773 (6)	0.246 (8)
ALMA-R 7B		0.813 (7)	0.759 (7)	0.233 (9)
Spearman's ρ		0.489	**0.508**	0.186

Table 4: Machine Translation models evaluated on neologisms with COMET, COMET-Kiwi, and BLEU. Rankings of models are provided for metrics and human evaluation for models used in §2. Spearman's ρ between each metric and human evaluation is also reported.

4 Key Findings

We utilize NEO-BENCH tasks to evaluate the ability of various LLMs to adapt to neologisms. The following are our key findings:

Current automatic metrics cannot accurately evaluate MT models that struggle with neologisms. In §2, MT models decrease in performance by 44% when translating neologisms with Bing being the best model based on human evaluation.

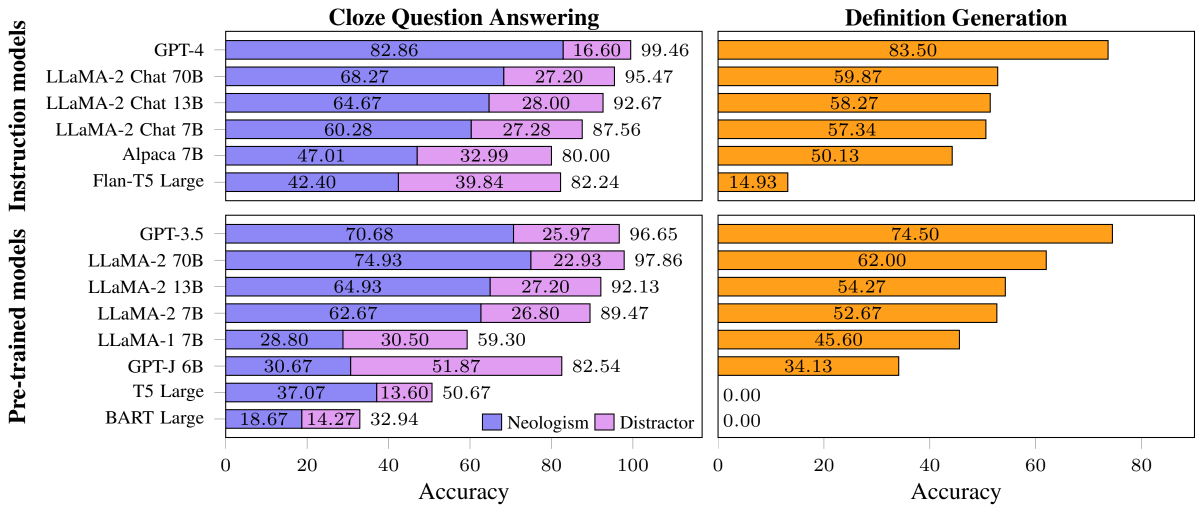

Figure 5: **Left**: Results of the Cloze Question Answering task reported by accuracy of selecting the neologism or distractor option. Combined accuracy for selecting either answer is provided. **Right**: Results of the Definition Generation task reported with accuracy of correct definitions. 5-shot prompting of models is used for both tasks.

However, COMET and COMET-Kiwi scores are notably high for all models as shown in in Table 4. The best models are Google Translate for COMET (0.854) and BLEU (0.507); and DeepL for COMET-Kiwi (0.807), highlighting that automatic metrics show poor system-level correlations with human judgments. For sentence-wise correlation between MT metrics and human evaluations, the average Spearman's ρ of COMET, COMET-Kiwi, and BLEU is 0.489, 0.508, and 0.186, respectively. In contrast, COMET-Kiwi, our highest correlating metric, has an average ρ of 0.629 for five language pairs on the WMT23 Quality Estimation task for direct assessment (Blain et al., 2023). From our reference sentences, translating neologisms often requires paraphrasing, resulting in low ρ for BLEU.

Models perform worse on neologisms than pre-existing words. For Cloze questions in Figure 5, neologism answers are designed to be more natural as the original passages contained these neologisms, yet all models select a large portion (27.5% on average) of distractor answers. Neologisms also have an average perplexity rank of 568 compared to distractor rankings of 47 in Figure 7, indicating much lower perplexity for pre-existing words.

Older LLMs perform significantly worse. In Figure 5, the average performance of GPT-J, BART, T5, and Flan-T5 is 30.31% lower in Cloze QA and 47.55% lower in Definition Generation than other models. In Figure 7, GPT-J and LLaMA-1 models exhibit higher neologism rankings than LLaMA-2 models, correlating with lower downstream performance. Newer models – GPT-3.5, GPT-4, LLaMA-2, and LLaMA-2 Chat – perform better as they are trained on data containing newer neologisms, generally have algorithmic improvements, and are trained with more resources than older models.

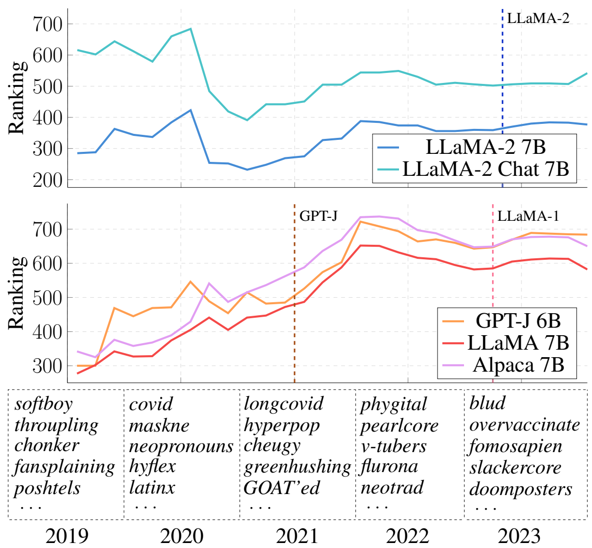

Figure 6: Rankings of neologisms over time compared to 5000 common words. LLaMA-2 models are plotted separately. Dashed lines show model knowledge cutoffs. Example neologisms from each year are provided, and neologisms without trendlines are reported at the end.

Perplexity rankings of older models increase drastically from 2019 until 2021. While neologisms are often gradually worked into a vocabulary (Zhu and Jurgens, 2021), we use trend lines to best estimate the date when a neologism becomes popular and report perplexity over time in Figure 6. LLaMA-2 rankings dip in 2020 but increase afterward as 52% of neologisms from this period are now conventionalized terms related to COVID-19.

Larger models handle neologisms better. In-

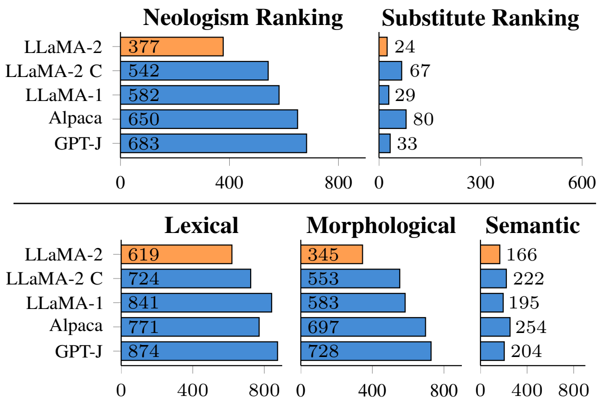

Figure 7: Average rankings of neologisms and pre-existing substitute terms compared to 5000 common words, sorted by model perplexities of texts filled in with each word. Neologisms are separated by linguistic type: lexical, morphological, and semantic.

creasing the sizes of LLaMa-2 and LLaMA-2 Chat leads to consistent improvements across both Cloze Question Answering and Definition Generation. On average, LaMA-2 70B and LLaMA-2 Chat 70B yield 10.13% higher accuracy in Cloze QA and 5.93% higher accuracy in Definition Generation than LLaMA-2 7B and LLaMA-2 Chat 7B.

Instruction-tuning results in high neologism perplexities. In Figure 7, LLaMA-1 and LLaMA-2 models have, on average, 116 lower neologism rankings than their instruct-tuned counterparts. Instruct models are trained with dialogue (Wei et al., 2022), so uncommon generation is less desired.

5 Linguistic Taxonomy Analysis

We separate NEO-BENCH task results by neologism linguistic structure: lexical, morphological, and semantic. Figure 7 presents perplexity rankings, Figure 9 reports human evaluation for MT, and Figure 10 shows the results for the best models on Cloze QA and Definition Generation.

Lexical neologisms produce the highest perplexities, but yield the best downstream results. On average, lexical neologisms have 185 higher rankings than other words, indicating higher relative model perplexities. Figure 8 shows the distribution of characters per token of neologisms using the LLaMA tokenizer, and the average number of characters per token for lexical, morphological, and semantic neologisms is 2.36, 2.98, and 3.24 respectively. Lexical neologisms have more fragmented tokenizations, as these words have the highest proportion of 1-2 character tokens. Lexical neologisms are less likely to be separated into long, common word roots or segments representing subword information, instead producing uncommon token sequences that result in higher neologism rankings and perplexity. In downstream tasks, however, lexical neologisms yield 0.3% higher Cloze accuracy, 7.1% more correct definitions, and 18.6% more good translations than other neologisms.

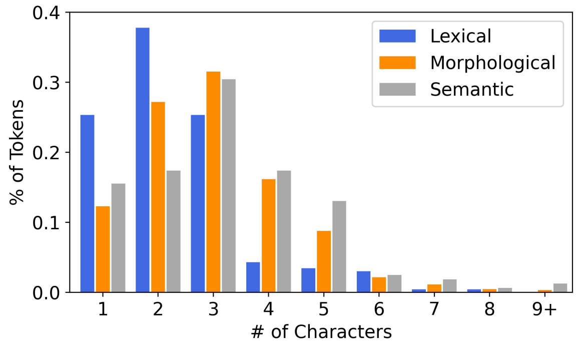

Figure 8: Distributions of characters per subword token of each neologism type, reported by the proportion of tokens that have a certain number of characters.

Morphological neologisms produce low perplexities but yield poor downstream performance. Compared to lexical neologisms, morphological neologisms are, on average, segmented into longer tokens and constructed with common subwords, resulting in lower perplexity rankings. However, they yield 4.0% lower Cloze accuracy, 7.5% more incorrect definitions, and 21.9% less good translations than lexical neologisms. 76.8% of neologisms without trend lines are morphological. Compared to lexical and semantic neologisms that require prevalence to be differentiated from incoherent strings, morphological neologisms are created with polynomial combinations of common subwords. Many of these intelligible combinations are largely unused, resulting in lower downstream performance.

Semantic neologisms produce the lowest perplexities and the worst performance in generation tasks. Since these neologisms use existing word forms, they have an average of 373 lower perplexity rankings than other neologism words. While semantic neologisms yield high Cloze QA accuracy, they also achieve the lowest percentages of correct definitions and translations. Models produce popular definitions and literal translations based on a word's most common meaning, as the new sense of semantic neologisms is often nuanced and difficult to capture.

6 Related Work

Temporal Drift in LLMs. Prior work has explored temporal data drift by creating temporal

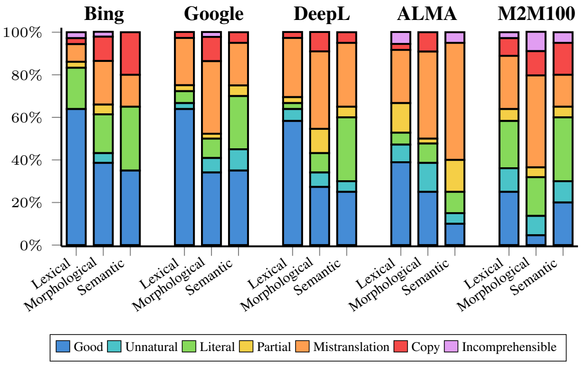

Figure 9: Results of the human-annotated MT models for each linguistic type of neologism.

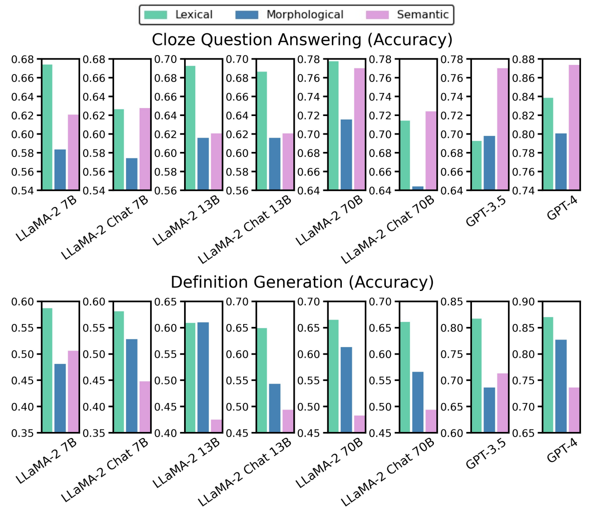

Figure 10: Results of Cloze QA and Definition Generation stratified by linguistic types of neologisms.

splits of training data data (Loureiro et al., 2022; Luu et al., 2022; Röttger and Pierrehumbert, 2021; Jin et al., 2022; Luu et al., 2022; Lazaridou et al., 2021). New factual updates of concepts are studied with temporal splits of text corpora and QA datasets (Jang et al., 2022; Margatina et al., 2023; Zhao et al., 2022; Vu et al., 2023). Other work has observed model degradation from new named entities (Onoe et al., 2022; Rijhwani and Preotiuc-Pietro, 2020; Chen et al., 2021). Temporal degradation occurs during short-term crisis events where information changes quickly (Pramanick et al., 2022). Studies have consistently found model degradation with perplexity and downstream tasks. There are no studies on model degradation from language change of neologisms, so we create a benchmark to evaluate models on neologisms with similar tasks.

Neologism Collection. Using reference texts as exclusion lists to filter common words from target corpora is the most documented method of neologism detection. Target texts and exclusion lists include news articles (Pinter et al., 2020; Falk et al., 2014), dictionaries (Kerremans et al., 2018; Langemets et al., 2020; Liu et al., 2013; Dhuliawala et al., 2016), social media (Pyo, 2023; Zalmout et al., 2019; Megerdoomian and Hadjarian, 2010) and other corpora (Cartier, 2017; Lejeune and Cartier, 2017). These texts are slow to curate, and semantic neologisms are filtered out. Moreover, no resource collects general semantic neologisms.

Some resources measure word prevalence with time-series data of search queries to collect single-word neologism candidates (Broad et al., 2018) or cybersecurity neologisms (Li et al., 2021). They are limited in scope by collecting only one type of neologism based on rising popularity. Other methods collect neologisms with new slang dictionary entries (Dhuliawala et al., 2016; Zhu and Jurgens, 2021). Dictionaries are slow to update, so new entries are often conventionalized words. There is no resource that uses time-series data to collect a variety of neologisms rising in prevalence.

Previous work has utilized search templates of explanation patterns to collect automatically neologisms (Breen et al., 2018). A few efforts have used neural methods to automatically detect one specific type of neologism, such as adjective-noun neologism pairs (McCrae, 2019), blend words (Megerdoomian and Hadjarian, 2010), and grammatical neologisms (Janssen, 2012; Falk et al., 2014), which are existing words with new parts of speech. No automatic resource collects neologisms from various topics and linguistic backgrounds. To address these limitations, NEO-BENCH uses multiple methods to semi-automatically collect a variety of neologisms, including multiword, semantic, and prevalent neologisms.

7 Conclusion

In this paper, we present NEO-BENCH, a new benchmark to test the ability of LLMs to generalize on neologisms. We use several methods to collect a variety of neologisms, including prevalent, multiword, and semantic neologisms. In our experiments, we find that models struggle with neologisms in both perplexity and downstream tasks. Machine Translation is especially difficult, as translating neologisms often requires paraphrasing the sentence. Current automatic metrics cannot measure translation quality, and human evaluation is still needed. Neologisms also affect models differently based on linguistic structure, indicating that this phenomenon is complex for LLMs to address.

Limitations

Most of our neologisms largely originated in US and UK English, as we collect textual data from news articles from this region. We do not restrict our locations for Reddit data, but the majority of English-speaking Reddit users are also from the same regions. Given our limited expertise in other English dialects, especially regions whose English variations are largely influenced by other languages, we do not collect many neologisms from English-speaking regions outside these regions. However, our computational framework for collecting neologisms can be applied to any language or local variation. For temporal drift of multilingual language modeling, we leave multilingual neologism collection and temporal drift analysis up for future work. Additionally, NEO-BENCH is static as we collect neologisms from mostly 2020-2023, which will become outdated over time. However, the semi-automatic collection methods require minimal human supervision and can be dynamically updated to continuously obtain neologisms. These methods require time-series data of words and online text corpora without needing human-curated information like updated dictionaries to filter words. The time-independent filters can collect recent neologisms without needing the time-consuming, manual curation of temporal splits of textual data.

Ethical Considerations

We utilize Reddit monthly dumps to obtain uncommon words, which often include sensitive information such as account usernames. We take the appropriate measures to ensure that no personally identifiable information (PII) is included in our dataset. We use a named-entity recognition model via SpaCy to identify named entities that are potentially PII and largely remove this information automatically when filtering for neologism candidates. We also manually inspect all the candidates to ensure that no PII is included in our dataset. As we use natural references from Google to construct our model inputs, we also review our hand-crafted sentences to ensure that there is no PII contained in these sentences. Many neologism entries in our work emerge from slang, and some slang words have expletive or offensive meanings. The purpose of our dataset and benchmark is to obtain a representative sample of neologisms and comprehensively evaluate the impact of neologisms on Large Language Models. We present examples that do not contain such offensive information, but these offensive entries are nonetheless a consistent source of neologisms. For expletive neologisms, we strive to create input sentences that capture the meaning of the neologism while not perpetuating gender, racial, and other potential biases. We do not collect any neologisms that are in direct reference to stereotypes and demographic biases.

References

Oshin Agarwal and Ani Nenkova. 2022. Temporal effects on pre-trained models for language processing tasks. *Transactions of the Association for Computational Linguistics*.

Ali Araabi, Christof Monz, and Vlad Niculae. 2022. How effective is byte pair encoding for out-of-vocabulary words in neural machine translation? In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 117–130, Orlando, USA. Association for Machine Translation in the Americas.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Frederic Blain, Chrysoula Zerva, Ricardo Ribeiro, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

James Breen, Timothy Baldwin, and Francis Bond. 2018. The company they keep: Extracting Japanese neologisms using language patterns. In *Proceedings of the 9th Global Wordnet Conference*, pages 163–171, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Claire Broad, Helen Langone, and David Guy Brizan. 2018. Candidate ranking for maintenance of an online dictionary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

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.

Lihu Chen, Gaël Varoquaux, and Fabian M. Suchanek. 2022. Imputing out-of-vocabulary embeddings with love makes language models robust with little cost.

Mingqing Chen, Rajiv Mathews, Tom Ouyang, and Françoise Beaufays. 2019. Federated learning of out-of-vocabulary words. *CoRR*, abs/1903.10635.

Shuguang Chen, Leonardo Neves, and Thamar Solorio. 2021. Mitigating temporal-drift: A simple approach to keep ner models crisp.

Nadezhda Chirkova and Sergey Troshin. 2021. A simple approach for handling out-of-vocabulary identifiers in deep learning for source code.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Shehzaad Dhuliawala, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. SlangNet: A WordNet like resource for English slang. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4329–4332, Portorož, Slovenia. European Language Resources Association (ELRA).

Ingrid Falk, Delphine Bernhard, and Christophe Gérard. 2014. From non word to new word: Automatically identifying neologisms in French newspapers. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4337–4344, Reykjavik, Iceland. European Language Resources Association (ELRA).

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Nobukazu Fukuda, Naoki Yoshinaga, and Masaru Kitsuregawa. 2020. Robust Backed-off Estimation of Out-of-Vocabulary Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4827–4838, Online. Association for Computational Linguistics.

Nicolas Garneau, Jean-Samuel Leboeuf, and Luc Lamontagne. 2018. Predicting and interpreting embeddings for out of vocabulary words in downstream tasks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 331–333, Brussels, Belgium. Association for Computational Linguistics.

David Heineman, Yao Dou, and Wei Xu. 2023. Thresh: A unified, customizable and deployable platform for fine-grained text evaluation.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Ziniu Hu, Ting Chen, Kai-Wei Chang, and Yizhou Sun. 2019. Few-shot representation learning for out-of-vocabulary words. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4102–4112, Florence, Italy. Association for Computational Linguistics.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maarten Janssen. 2012. NeoTag: a POS tagger for grammatical neologism detection. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2118–2124, Istanbul, Turkey. European Language Resources Association (ELRA).

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. 2022. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 1–16, virtual+Dublin. Association for Computational Linguistics.

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. *Pragmatics and Cognition*, 25(1):174–200.

Margit Langemets, Jelena Kallas, Kaisa Norak, and Indrek Hein. 2020. New estonian words and senses: Detection and description. *Dictionaries: Journal of the Dictionary Society of North America*, 41(1):69–82.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021.

Mind the gap: Assessing temporal generalization in neural language models.

Gaël Lejeune and Emmanuel Cartier. 2017. Character based pattern mining for neology detection. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 25–30, Copenhagen, Denmark. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Ying Li, Jiaxing Cheng, Cheng Huang, Zhouguo Chen, and Weina Niu. 2021. NEDetector: Automatically extracting cybersecurity neologisms from hacker forums. *Journal of Information Security and Applications*, 58:102784.

Shuheng Liu and Alan Ritter. 2023. Do CoNLL-2003 named entity taggers still work well in 2023? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Tsun-Jui Liu, Shu-Kai Hsieh, and Laurent Prévot. 2013. Observing features of PTT neologisms: A corpus-driven study with n-gram model. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing, ROCLING 2015, National Sun Yat-sen University, Kaohsiung, Taiwan, October 4-5, 2013*. Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taiwan.

Johannes V. Lochter, Renato M. Silva, and Tiago A. Almeida. 2022. Multi-level out-of-vocabulary words handling approach. *Knowledge-Based Systems*, 251:108911.

Johannes V. Lochter, Renato Moraes Silva, and Tiago A. Almeida. 2020. Deep learning models for representing out-of-vocabulary words. *CoRR*, abs/2007.07318.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States. Association for Computational Linguistics.

Katerina Margatina, Shuai Wang, Yogarshi Vyas, Neha Anna John, Yassine Benajiba, and Miguel Ballesteros. 2023. Dynamic benchmarking of masked language models on temporal concept drift with multiple views. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2881–2898, Dubrovnik, Croatia. Association for Computational Linguistics.

John Philip McCrae. 2019. Identification of adjective-noun neologisms using pretrained language models. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 135–141, Florence, Italy. Association for Computational Linguistics.

Karine Megerdoomian and Ali Hadjarian. 2010. Mining and classification of neologisms in persian blogs. In *HLT-NAACL 2010*.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spaCy: v3.7.2: Fixes for APIs and requirements.

Anmol Nayak, Hariprasad Timmapathini, Karthikeyan Ponnalagu, and Vijendran Gopalan Venkoparao. 2020. Domain adaptation challenges of BERT in tokenization and sub-word representations of out-of-vocabulary words. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 1–5, Online. Association for Computational Linguistics.

Yasumasa Onoe, Michael Zhang, Eunsol Choi, and Greg Durrett. 2022. Entity cloze by date: What LMs know about unseen entities. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 693–702, Seattle, United States. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yuval Pinter, Cassandra L. Jacobs, and Max Bittker. 2020. NYTWIT: A dataset of novel words in the New York Times. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6509–6515, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Aniket Pramanick, Tilman Beck, Kevin Stowe, and Iryna Gurevych. 2022. The challenges of temporal alignment on Twitter during crises. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2658–2672, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiyoon Pyo. 2023. Detection and replacement of neologisms for translation. Master's thesis, The Cooper Union for the Advancement of Science and Art.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shruti Rijhwani and Daniel Preotiuc-Pietro. 2020. Temporally-informed analysis of named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Paul Röttger and Janet Pierrehumbert. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David Mortensen, and Yulia Tsvetkov. 2020. Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 367–376, New York, New York. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2019. Neural machine translation with byte-level subwords. *CoRR*, abs/1909.03341.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models.

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.

Zhixue Zhao, George Chrysostomou, Kalina Bontcheva, and Nikolaos Aletras. 2022. On the impact of temporal concept drift on model explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4039–4054, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jian Zhu and David Jurgens. 2021. The structure of online social networks modulates the rate of lexical change. *CoRR*, abs/2104.05010.

A Other Related Work

Unseen Words. Rare words or typos are typically unseen in data when training a model but may show up during inference. Prior work has measured model degradation from unseen words (Chirkova and Troshin, 2021; Nayak et al., 2020) and used contextual subword embeddings (Garneau et al., 2018; Chen et al., 2019; Hu et al., 2019; Wang et al., 2019; Araabi et al., 2022), similar surface forms of common words (Chen et al., 2022; Fukuda et al., 2020), and morphological structure (Lochter et al., 2020, 2022) to represent unseen words. Comparatively, the neologism lifecycle follows three stages: emergence, dissemination, and conventionalization (Cartier, 2017). New words often become prevalent and drastically shift a language's distribution. Semantic neologisms also use existing word forms and are not classified as unseen words.

Table 5 provides an overview and comparison of English neologism resources, including the types of neologisms and collection methods used in each dataset. NEO-BENCH uses more methods and collects more types of neologisms.

B Data Collection

For Reddit neologism candidates, we collected 500 million utterances in December 2021 and 200 million utterances from January to May 2022. We tokenize the utterances with the NLTK package (Bird et al., 2009) to get individual word counts and update a generic word counter. Neologism candidates are selected by filtering out typos and extremely rare words with less than a frequency of 50. We further filter out named entities by utilizing a SpaCy named entity recognition (NER) model (Honnibal and Montani, 2017) (`en_core_web_sm`) to detect proper nouns and update a named entity counter. We compare the counts of words from the general counter and the named entity counter and filter out the word if the proportion that a general word is in a named entity is greater than 0.5. The remaining words with the lowest frequencies are the list of uncommon words that we treat as neologism candidates for a given month. In total, we collect 74,542 neologism candidates.

For news articles, we use a script to collect 11,412 headlines from Google News from 2019-2023. In total, we get 60,671 noun and verb phrases with a Part-of-Speech Tagger via SpaCy (`en_core_web_sm`) that we treat as neologism candidates. We use an old dataset of 80,071 neologisms obtained from two slang dictionaries (Zhu and Jurgens, 2021) and sample 200 neologisms with interesting or no trend lines. Table 6 provides the breakdown of method overlap between each method pair in NEO-BENCH. Instead of the sample of 1,100 data points, we compare a total of all 6,908 words tweeted out by the NYT First Said bot from 2020 to 2023 with the other methods. Figure 11 provides the breakdown of NEO-BENCH by collection method and linguistic type.

B.1 Google Trends Filtering

We collect Google Trends monthly data from January 2010 to July 2023. While Google Trends provides data from 2004, there are inconsistencies in word usage frequencies until 2010. To compare word prevalence between neologisms, we do a pairwise comparison of a neologism candidate with the misspelling 'dangrous', which provides a consistent baseline comparison for word usage data. We then use this normalized trend line for neologism candidate filtering.

In total, we create five differing methods that use a combination of filtering criteria, including curve-fitting, argmax detection, integral, line of best-fit, and maximum trend data values, to evaluate words as neologism candidates. We select the best combination based on which method yields both high precision and estimated recall in collecting neologisms. Using 20,000 words collected in February 2022, we filter this set through all five methods which filter out almost 90% of words. We sample each method for 100 candidates and manually annotate the samples for neologism classification, obtaining a sampled precision of each method. We combine all the neologisms from the manually-annotated samples to obtain a computationally derived neologism set. We evaluate each method for its estimated recall based on the proportion of words from the computational neologism sample that is not filtered out. The sample precision is particularly low given the sparsity of neologisms that appear at a specific point in time, so we select the method with the highest precision of 0.2 and an estimated recall of 0.625 to reduce the amount of manual annotation required.

B.2 Dataset Analysis

With the best-performing filter method, we also estimate its recall with a set of 100 handpicked neologisms in our dataset. The estimated recall of our method is 0.55. This estimate is slightly

Dataset	Neologism Type: Emerging?	Multiword?	Semantic?	Generalized?	Collection Method: Exclusion Lists	Dictionaries	Time-Series	Templates
(Pinter et al., 2020)	✗	✗	✗	✓	✓	✗	✗	✗
(Kerremans et al., 2018)	✗	✗	✗	✓	✓	✗	✗	✗
(Zalmout et al., 2019)	✗	✗	✗	✓	✓	✗	✗	✗
(Janssen, 2012)	✗	✗	✓	✗	✓	✗	✗	✗
(Dhuliawala et al., 2016)	✗	✓	✗	✗	✗	✓	✗	✗
(Zhu and Jurgens, 2021)	✗	✓	✗	✓	✗	✓	✗	✗
(McCrae, 2019)	✗	✓	✗	✓	✗	✓	✗	✗
(Broad et al., 2018)	✓	✗	✗	✓	✗	✗	✓	✗
(Li et al., 2021)	✗	✓	✗	✗	✗	✗	✓	✗
NEO-BENCH (this work)	✓	✓	✓	✓	✓	✓	✓	✓

Table 5: Comparison of English Neologism resources by the types of neologisms collected and the collection method used. NEO-BENCH covers more types of neologisms by using more methods than prior neologism datasets.

Source **# Candidates**	**Reddit** 74542	**News** 60671	**NYTimes** 6908	**Dictionary** 80071
% Reddit	-	0.93%	0.94%	3.91%
% News	0.76%	-	0.26%	0.12%
% NYTimes	0.09%	0.03%	-	0.15%
% Dictionary	4.19%	0.16%	1.80%	-
% Total	5.04%	1.12%	3.00%	4.18 %

Table 6: Number of shared neologism candidates for each method pair. The overlap is reported as a percent of the total number of candidates for each method.

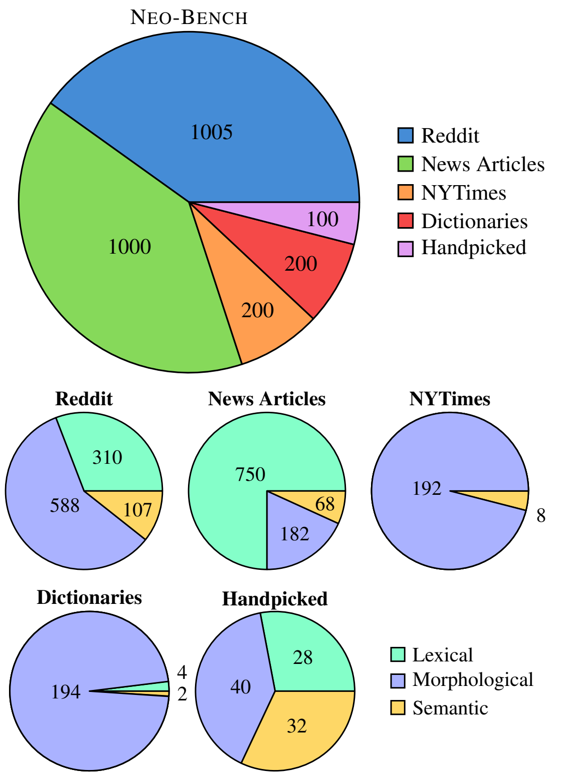

Figure 11: Breakdown of NEO-BENCH by collection method. Each method is further stratified by the linguistic type of neologisms.

lower than the estimated recall we used with the neologism candidates we computationally gathered, but the method remains consistent.

Analyzing the overlap of words in our dataset with Urban Dictionary by collection method, we find that 44.37% of Reddit neologisms, 38.4% of News neologisms, and 65.25% of neologisms obtained from dictionaries and exclusion lists overlap.

B.3 Emergence Date Labeling for Perplexity-Based Rankings

Using Google Trends to record the month and year where word usage spikes, we estimate the date for when a neologism enters the dissemination stage of its lifecycle. We find that 68% of neologisms emerged during 2020-2023, and 17% of all words have no dissemination date or trend line. The remaining words were prevalent before 2020, but a new connotation or usage has recently emerged. These dates are potentially inaccurate as a trend line is collapsed into a single date. Compared to using the entire graph to evaluate long-term word trends when filtering for neologisms, a single date may not perfectly capture neologism growth for words that exhibit a steady rise in growth.

B.4 NEO-BENCH Tasks

We use a script to collect Google Search Descriptions as references of natural occurrences of neologisms to construct model inputs. We work with in-house native English speakers to construct 100 minimal pair sentences, 240 sentences containing neologisms, 750 natural questions, and 750 Cloze passages with neologism and distractor answers. We also work with in-house native Chinese speakers to construct 240 reference translations for automatic MT evaluation. All of the annotators have a college-level education.

C Experimental Details

All models are evaluated on two NVIDIA A40 GPUs for a single run since models are not fine-tuned for NEO-BENCH tasks.

C.1 Machine Translation

Table 10 provides example outputs for each translation category. We provide a correct human translation for outputs.

We did not use MBART as we found 24 instances of English text in MT outputs. MBART copied longer English sequences compared to other models, and there was 1 instance where it produced non-Chinese output.

For human evaluation, we crowdsource annotations from 5 native Chinese speakers from Prolific[5] residing in the United States and United Kingdom. All of the annotators have a college education and are informed about the nature of the study. Each annotator was given the same set of neologism sentences across all 5 models evaluated to ensure a standard comparison between models. The average time to annotate 20 minimal pairs was 80 minutes, and each annotator was paid $12.00 an hour, which is on the high end for standard pay on Prolific. We use the Thresh (Heineman et al., 2023) interface for annotating translation sentences, and Figure 13 provides a screenshot of the interface.

C.2 Rank Classification with Perplexity

We also tested T5 and Flan-T5 for perplexity ranking and find that Flan-T5 exhibits higher neologism rankings than T5. However, when sorting words by lowest perplexity and filling in the mask, we find that these models produced entirely incoherent sequences, so we do not report these models. Given the computational intensity of evaluating 5,002 sequence perplexities, we only evaluate the base size of models.

C.3 Cloze Question Answering

Rank classification is used to select the lowest perplexity answer for BART, T5, and GPT-J. We shuffle the order of answers and conduct experiments with 5-shot prompting with the following format:

```
Fill in the blank with the options below:
Question: [EXAMPLE CLOZE PASSAGE]
a) [EXAMPLE INCORRECT ANSWER]
b) [EXAMPLE DISTRACTOR ANSWER]
c) [EXAMPLE INCORRECT ANSWER]
d) [EXAMPLE INCORRECT ANSWER]
e) [EXAMPLE NEOLOGISM ANSWER]
f) [EXAMPLE INCORRECT ANSWER]
Answer: e) [EXAMPLE NEOLOGISM ANSWER]
...
Fill in the blank with the options below:
Question: [TEST CLOZE PASSAGE]
a) [TEST INCORRECT ANSWER]
b) [TEST INCORRECT ANSWER]
c) [TEST NEOLOGISM ANSWER]
d) [TEST DISTRACTOR ANSWER]
e) [EXAMPLE INCORRECT ANSWER]
f) [EXAMPLE INCORRECT ANSWER]
Answer:
```

[5] https://www.prolific.com

C.4 Definition Generation

We conduct experiments with 5-shot prompting with the following prompt:

```
Answer the question.
Question: [EXAMPLE DEFINITION QUESTION]
Answer: [EXAMPLE DEFINITION ANSWER]
...
Answer the question.
Question: [TEST CLOZE PASSAGE]
Question: [TEST DEFINITION QUESTION]
Answer:
```

One of the paper's authors manually annotates 100 outputs. We measure the Cohen's Kappa between automatic GPT-4 and human evaluation, and we obtain an aggregate Cohen's Kappa of 0.778, indicating high agreement between human judgment and GPT-4.

For automatic evaluation, we additionally determine if a correct model definition is better or worse than the reference definition provided by human input. For incorrect answers, we separate between incorrect and omitted generations, which are model outputs that are either left blank or, for GPT models, outputs where the model acknowledges that it does not recognize the neologism.

Table 11 provides the full results of the open-domain question-answering experiments, including the average length of definitions, manual evaluation and model-wise Cohen's Kappa. While LLaMA-2 70B outperforms GPT-3.5 in Cloze QA, GPT 3.5 produces more correct definitions than LLaMA-2-70B. Instruction-tuned models produce a higher proportion of correct answers that are deemed better than the human reference sentence. We report the average length of generations for each model and conclude that GPT-4 prefers instruction model outputs because the human reference

Model	Pre-trained Lexical	Morphological	Semantic
BART-Large	19.88	18.48	14.94
T5-Large	33.54	39.88	39.08
GPTJ 6B	34.78	25.22	36.78
LLaMA-1 7B	29.50	27.86	29.89
LLaMA-2 7B	67.39	58.36	62.07
LLaMA-2 13B	69.25	61.58	62.07
LLaMA-2 70B	**77.95**	**71.55**	77.01
GPT 3.5	69.25	69.79	77.01
Model	**Instruction-Tuned Lexical**	**Morphological**	**Semantic**
Flan-T5 Large	41.99	42.99	41.61
Alpaca 7B	50.37	44.87	42.99
LLaMA-2 Chat 7B	62.64	57.42	62.76
LLaMA-2 Chat 13B	68.63	61.58	62.07
LLaMA-2 Chat 70B	71.43	64.22	72.41
GPT-4	**83.85**	**80.06**	**87.36**

Table 7: Neologism accuracies of models for the Cloze Question Answering task separated by linguistic type: lexical (322), morphological (341), and semantic (87). Best performing accuracy is presented in **bold**, while highest shared accuracy is reported in underline.

sentences are on average 19.20 words long, which is more concise compared to the more elaborative responses of instruct-tuned models. We find that 82.27% of preferred answers across all models are longer than the alternative correct answer. Table 12 provides examples of instruct-tuned model responses that are evaluated as better than the human reference and examples of GPT-omitted responses. Even when prompted with shortened answers, instruction-tuned models produce longer-form responses. We did not test for restraining the length of the model output as the complexity and reference definition length for each neologism varies extensively. Instruction-tuned models are less likely to produce omitted answers, and only GPT-3.5 and GPT-4 have generations that acknowledge that a term is unrecognized.

D Linguistic Taxonomy

Table 7 provides the linguistic breakdown of neologism accuracies for each model in Cloze QA. Table 9 provides the stratified results of definition generation by linguistic type of neologism. Table 8 provides the breakdown of automatic machine translation evaluation, and Figure 12 provides the stratified results of manually labeling translations by linguistic type. We report each category as a proportion to the total amount of neologisms for a certain linguistic type. ALMA also produces the highest rate of mistranslations across all models when translating semantic neologisms. Model performance discrepancy between the open source models and commercial systems is highest for lexical neologisms. For automatic metrics, lexical neologisms do not yield higher BLEU scores but yield higher COMET and COMET-Kiwi scores. Morphological neologisms yield 4.3 lower BLEU scores, 2.7% lower COMET scores, and 4.1% lower COMET-Kiwi scores than lexical neologisms. BLEU scores for semantic neologisms are high as there is high token overlap between different senses of the same word form but often yield similarly low COMET and COMET-Kiwi scores as Morphological neologisms.

Model	COMET			COMET-Kiwi			BLEU		
	Lexical	Morphological	Semantic	Lexical	Morphological	Semantic	Lexical	Morphological	Semantic
Google Translate	**0.870**	**0.842**	**0.849**	0.820	0.782	0.805	**0.530**	**0.487**	**0.507**
Bing Translator	0.852	0.806	0.812	0.812	0.769	0.786	0.484	0.418	0.467
DeepL Translator	0.856	0.833	0.833	**0.823**	**0.792**	**0.814**	0.434	0.373	0.429
GPT-4	0.856	0.836	**0.849**	0.796	0.772	0.772	0.458	0.410	0.486
GPT-3.5	0.850	0.831	0.817	0.809	0.769	0.765	0.409	0.360	0.421
ALMA-R 13B	0.818	0.824	0.801	0.795	0.759	0.762	0.248	0.246	0.240
ALMA-R 7B	0.827	0.805	0.802	0.784	0.741	0.750	0.244	0.215	0.256
ALMA-7 B	0.814	0.795	0.786	0.770	0.731	0.730	0.303	0.262	0.287
M2M100 1.2B	0.816	0.743	0.774	0.786	0.715	0.730	0.357	0.310	0.361

Table 8: COMET and BLEU scores of Machine Translation models when evaluated on sentences containing neologisms. Model performance is separated by linguistic type of neologisms and aggregate score.

Pre-trained Models	Lexical (322)			Morphological (341)			Semantic (87)		
	Correct	% Worse	% Better	Correct	% Worse	% Better	Correct	% Worse	% Better
GPTJ 6B	0.367	78.7%	21.3%	0.323	68.1%	31.9%	0.322	78.6%	21.4%
LLaMA-1 7B	0.506	72.9%	27.1%	0.437	68.4%	31.6%	0.345	63.2%	36.8%
LLaMA-2 7B	0.587	65.1%	34.9%	0.481	58.0%	42.0%	0.506	72.7%	27.3%
LLaMA-2 13B	0.609	62.7%	37.3%	0.510	62.2%	37.8%	0.425	64.9%	35.1%
LLaMA-2 70B	0.665	57.4%	42.6%	0.613	54.0%	46.0%	0.483	57.1%	42.9%
GPT 3.5	0.817	6.1%	93.9%	0.686	9.8%	90.2%	0.713	11.4%	88.6%
Instruct-tuned Models	**Lexical (322)**			**Morphological (341)**			**Semantic (87)**		
	Correct	% Worse	% Better	Correct	% Worse	% Better	Correct	% Worse	% Better
Flan-T5 Large	0.158	96.2%	3.8%	0.158	83.5%	16.5%	0.080	86.3%	13.7%
Alpaca 7B	0.565	71.0%	29.0%	0.463	70.8%	29.2%	0.414	63.8%	36.2%
LLaMA-2 Chat 7B	0.581	59.4%	40.6%	0.528	63.3%	36.7%	0.448	66.7%	33.3%
LLaMA-2 Chat 13B	0.649	40.7%	59.3%	0.543	41.1%	58.9%	0.494	39.5%	60.5%
LLaMA-2 Chat 70B	0.661	46.0%	54.0%	0.566	44.5%	55.5%	0.494	37.2%	62.8%
GPT-4	0.870	8.3%	91.7%	0.827	11.4%	88.6%	0.736	11.0%	89.0%

Table 9: Results of the definition generation task when separated by linguistic type. Accuracy is reported as a proportion of correct answers compared to the total number of neologisms of each linguistic type. The percentages of correct answers that are labeled as 'worse' and 'better' than the human reference sentence by GPT-4 are provided.

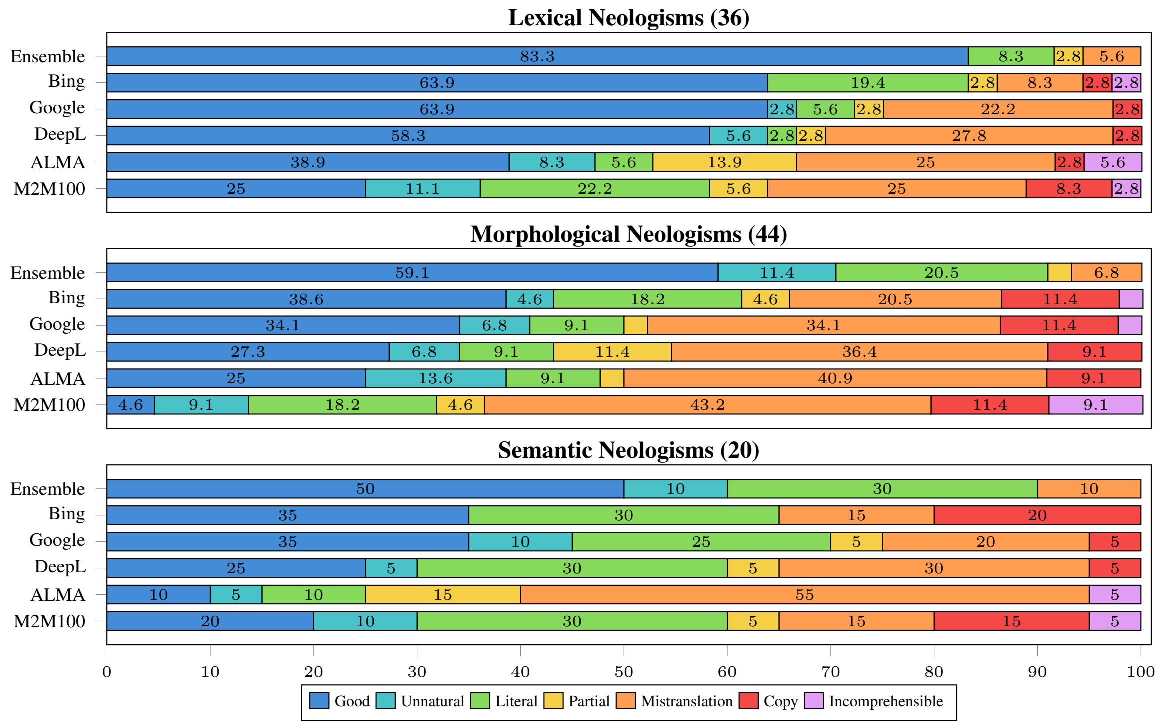

Figure 12: Results of the Machine Translation task with human-annotated labels for each linguistic type of neologism. Results are reported as percentages of the total number of neologisms of each linguistic category (provided in the titles). A Human Oracle Ensemble selecting the best model translation for each sentence is provided.

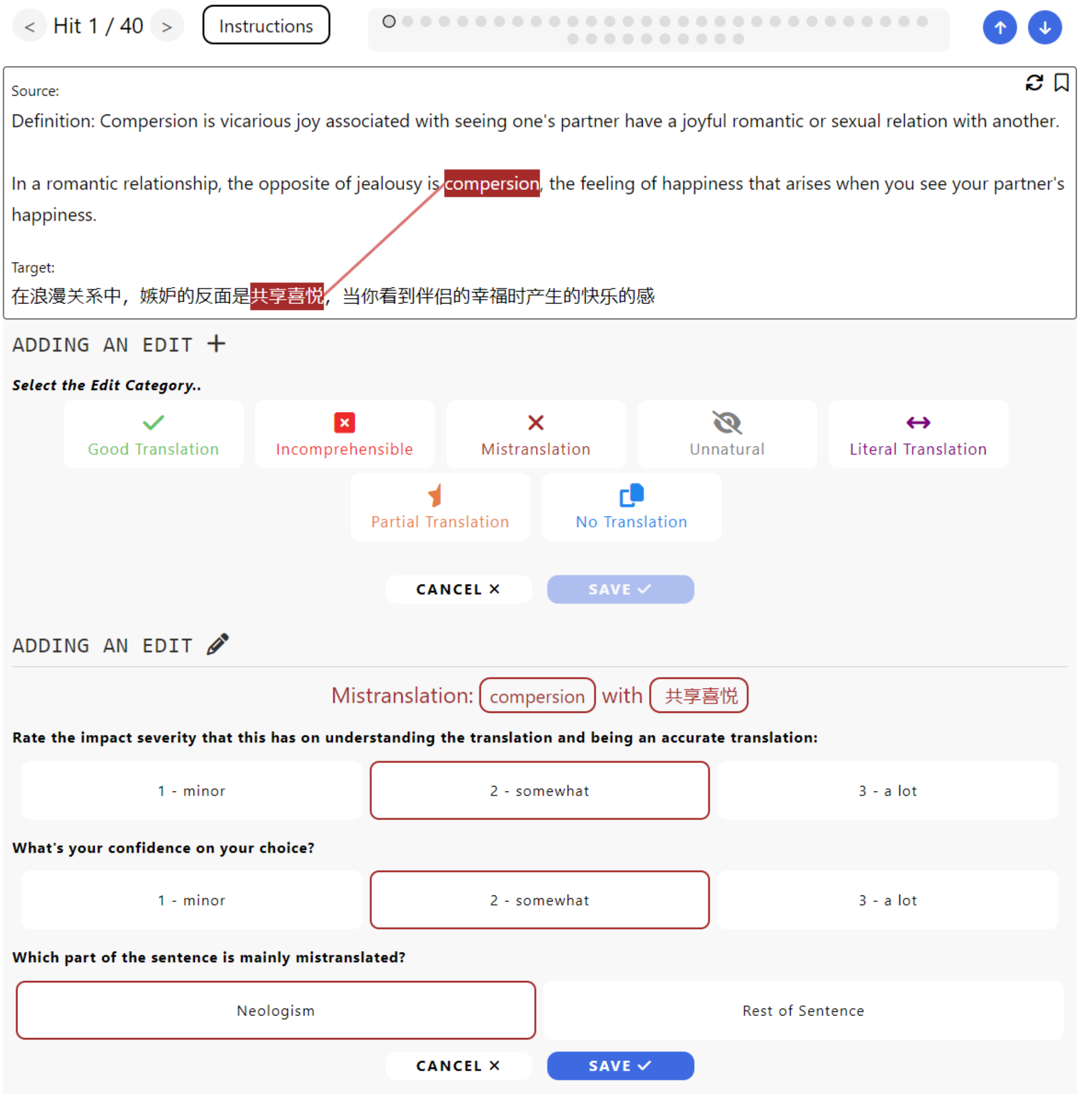

Figure 13: Thresh Interface used to crowdsource human annotations of Machine Translation Output.

Translation Output Examples	
Fire Weather	**Input:** They will not issue official warnings until **fire weather** is forecast to occur. *(Fire weather is the use of meteorological parameters such as relative humidity, wind speed, mixing heights, and soil moisture to determine whether conditions are favorable for fire growth and smoke dispersion.)*
	Model Output (Good): 在预测发生火灾天气之前，他们不会发布官方警告。 *(Before predicting the occurrence of* ***fire weather****, they will not issue an official warning.)*
	Human Translation: 他们直到预测到火灾天气才会发布官方警告。 *(They will only issue official warnings when a* ***fire weather*** *is forecasted.)*
Dupe	**Input:** Discover new affordable **dupes** for luxury expensive makeup products. *(A dupe is an abbreviation of the word "duplicate".)*
	Model Output (Unnatural): 为奢华昂贵的彩妆产品发现新的负担得起的复制品。 *(Find new affordable* ***replicas*** *for luxurious and expensive makeup products.)*
	Human Translation: 探索奢侈昂贵化妆品的平价替代新品。 *(Explore new affordable* ***alternatives*** *for luxurious and expensive makeup products.)*
Snowvember	**Input:** When sleet started falling during Thanksgiving it was officially **snowvember**. *(Snowvember refers to a particular November that experiences a lot of snowfall.)*
	Model Output (Literal): 当感恩节期间开始下雨夹雪时，正式进入了十一月雪。 *(When it starts to rain and snow during Thanksgiving, it officially enters the* ***November snow****.)*
	Human Translation: 感恩节开始下的雨夹雪标志着‘雪月’的正式开始。 *(The start of sleet during Thanksgiving marks the official beginning of* ***'Snowvember'.****)*
Trollbaiting	**Input:** **Trollbaiting** has caused my growth on social media this past month to be super high. *(Trollbaiting describes when an internet user knowingly invites the hatred of a highly reactionary group of trolls.)*
	Model Output (Partial): 在过去的一个月里，我在社交媒体上的增长速度超快。 *(In the past month, my growth on social media has been extremely fast.)*
	Human Translation: 过去一个月里，通过挑衅网络喷子，我在社交媒体上的增长极为迅速。 *(In the past month, by* ***provoking internet trolls****, my growth on social media has been extremely rapid.)*
Forever Chemicals	**Input:** The environment cannot break down **forever chemicals**, and they will remain in our bodies for years if ingested. *(Forever chemicals are used to make products grease-proof, water-proof, stick-proof, and stain-resistant and are toxic to humans and nearly indestructible.)*
	Model Output (Mistranslation): 环境不能永远分解化学物质，如果摄入它们，它们会在我们体内停留数年。 *(The environment cannot* ***always break down chemicals****, and if ingested, they can stay in our bodies for years.)*
	Human Translation: 环境无法分解永久化学物质，一旦摄入，这些物质将在我们的身体中残留多年。 *(The environment cannot break down* ***'forever chemicals',*** *and once ingested, these substances will remain in our bodies for many years.)*
Blud	**Input:** What is **blud** talking about I can't understand. *(Blud is slang that is used to address men and means bro.)*
	Model Output (Copy): **blud** 在说什么我听不懂。 *(****blud****, I don't understand what you are saying.)*
	Human Translation: 我听不懂这哥们在说什么。 *(I can't understand what this* ***guy*** *is saying.)*
Noctor	**Input:** Is the narcissist in your life a **noctor** and diagnosing you? (A noctor is a health professional (usually nurse) who takes on some traditional roles performed by the doctor.)
	Model Output (Incomprehensible): 你生活中的自恋者是夜幕降临，诊断着你吗? *(Is the narcissist in your life like* ***nightfall****, diagnosing you?)*
	Human Translation: 那个在你生活中的自恋者是不是冒充医生给你做诊断? *(Is the narcissist in your life* ***pretending to be a doctor*** *and diagnosing you?)*

Table 10: Example model outputs for all possible translation categories. For each neologism example, the English input and Chinese output is reported. A gold reference definition of the neologism is provided. (Neologism definitions and English translations are shown for information only.)

	Model	Incorrect	Omitted	Worse	Better	Avg. Length	Acc. (↑)	Human Eval. (100) Acc. (↑)	Cohen's κ (↑)
		GPT-4 Eval. (750)							
Pre-trained	GPT-J	494	0	190	66	19.04	0.341	0.38	0.711
	LLaMA-1	384	24	240	102	16.69	0.456	0.48	0.697
	LLaMA-2	311	42	251	144	19.49	0.529	0.61	0.681
	LLaMA-2 13B	262	81	255	152	18.76	0.544	0.56	0.698
	LLaMA-2 70B	191	94	260	205	17.29	0.620	0.67	0.827
	GPT 3.5	95	96	**46**	513	41.21	0.745	0.72	0.828
Instruct	Flan-T5	638	0	100	12	15.42	0.149	0.17	0.670
	Alpaca	374	0	112	264	34.18	0.501	0.56	0.761
	LLaMA-2 Chat	320	0	155	275	24.02	0.573	0.58	0.758
	LLaMA-2 Chat 13B	313	0	178	259	24.14	0.583	0.66	0.731
	LLaMA-2 Chat 70B	300	1	200	249	24.22	0.599	0.64	0.771
	GPT 4	bf106	18	62	**564**	38.09	**0.835**	**0.85**	**0.891**

Table 11: Full results of the Definition Generation task showing the number of **correct** and **incorrect** answers per model. A sample of 100 neologisms are manually evaluated, and Cohen's Kappa is calculated to determine annotator agreement between GPT-4 and human evaluation. Model accuracy is reported for both manual and GPT-4 evaluation based on the number of definitions evaluated. Correct outputs are divided between being worse or better than the human reference sentences. Wrong definitions are divided between incorrect or omitted output where models do not provide a definition when prompted. Average length of definitions is reported with number of words.

Term	Definitions
a) Better Model Output Examples	
Inclusive Education	**Reference Definition:** *Inclusive education refers to all students being able to access and gain equal opportunities to education and learning.*
	Flan – T5 Large Output: *Inclusive education is a philosophy that emphasizes the inclusion of all students, regardless of their disability, gender, race, religion, or sexuality.*
Greenwashing	**Reference Definition:** *Greenwashing is the act or practice of making a product, policy, activity, etc. appear to be more environmentally friendly or less environmentally damaging than it really is.*
	Alpaca 7B Output: *Greenwashing is the practice of misleading consumers about the environmental practices of a company or product. It involves claiming to be environmentally friendly without having any genuine commitment to sustainability.*
TCG	**Reference Definition:** *TCG is an acronym for "trading card game."*
	LLaMA – 2 Chat 7B Output: *TCG stands for Trading Card Game, a type of game where players use cards to represent characters, spells, and other elements to battle each other.*
Mermaidcore	**Reference Definition:** *Mermaidcore is an aesthetic focusing on imagery and fashion related to mermaids.*
	LLaMA – 2 Chat 13B Output: *Mermaidcore is a type of aesthetic that features pastel colors, sparkly accessories, and flowing fabrics, inspired by the mythical creatures known as mermaids.*
Wokeness	**Reference Definition:** *Wokeness is the quality of being alert and concerned about social injustice and discrimination.*
	LLaMA – 2 Chat 70B Output: *It is a term used to describe the state of being aware of and attuned to social injustices, particularly those related to racism, sexism, and economic inequality.*
Longcovid	**Reference Definition:** *Longcovid is the signs and conditions that continue or develop after acute COVID-19 infection.*
	GPT – 4 Output: *Long COVID, also known as post-acute sequelae of SARS-CoV-2 infection (PASC), refers to a range of symptoms that can continue for weeks or months after the acute phase of a COVID-19 infection has resolved. Symptoms may include fatigue, breathlessness, brain fog, depression and anxiety, pain, and other long-term complications.*
b) Omitted Model Output Examples	
Falstalgia	**Reference Definition:** *Falstalgia is a misguided feeling of nostalgia, a longing for a past that really wasn't that great and you weren't around for.*
	GPT – 4 Output: *I'm sorry, "falstalgia" does not appear to be a recognized term in English language or any other languages. It could be a typo. If you meant "nostalgia," it is a sentimental longing or wistful affection for the past, typically for a period or place with happy personal associations.*
Flurona	**Reference Definition:** *Flurona is a coinfection with influenza and COVID-19.*
	GPT – 3.5 Output: *I'm sorry, but I couldn't find any information on the term "flurona." It is possible that it is a newly coined term or a term that is not widely recognized.*

Table 12: Example model outputs of the definition generation task. **a)** Instruction-tuned model outputs evaluated as better than the reference definition by GPT-4 and **b)** GPT model outputs that omit definitions are provided.