Computational Linguistics in the Netherlands Journal 13 (2024) 205-232 Submitted 02/2024; Published 03/2024

Beyond Perplexity: Examining Temporal Generalization in Large Language Models via Definition Generation

Iris Luden[*] IRISLUDEN@GMAIL.COM
Mario Giulianelli[**] MGIULIANELLI@INF.ETHZ.CH
Raquel Fernández[*] RAQUEL.FERNANDEZ@UVA.NL

[*] *Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands*

[**] *Department of Computer Science, ETH Zürich, Switzerland*

Abstract

The advent of large language models (LLMs) has significantly improved performance across various Natural Language Processing tasks. However, the performance of LLMs has been shown to deteriorate over time, indicating a lack of temporal generalization. To date, performance deterioration of LLMs is primarily attributed to the factual changes in the real world over time. However, not only the facts of the world, but also the language we use to describe it constantly changes. Recent studies have indicated a relationship between performance deterioration and semantic change. This is typically measured using perplexity scores and relative performance on downstream tasks. Yet, perplexity and accuracy do not explain the effects of temporally shifted data on LLMs in practice.

In this work, we propose to assess lexico-semantic temporal generalization of a language model by exploiting the task of contextualized word definition generation. This in-depth semantic assessment enables interpretable insights into the possible mistakes a model may perpetrate due to meaning shift, and can be used to complement more coarse-grained measures like perplexity scores. To assess how semantic change impacts performance, we design the task by differentiating between semantically stable, changing, and emerging target words, and experiment with `T5-base`, fine-tuned for contextualized definition generation.

Our results indicate that (i) the model's performance deteriorates for the task of contextualized word definition generation, (ii) the performance deteriorates more for semantically changing words compared to semantically stable words, (iii) the model exhibits significantly lower performance and potential bias for emerging words, and (iv) the performance does not correlate with cross-entropy or (pseudo)-perplexity scores.[1] Overall, our results show that definition generation can be a promising task to assess a model's capacity for temporal generalization with respect to semantic change.

1. Introduction

Large language models (LLMs) are increasingly employed for linguistic tasks that require advanced language understanding and generation such as question answering, text summarization, and machine translation. These models are typically pre-trained on large text corpora, from which they learn the intricate patterns of human natural language use. Despite their overall impressive performance, LLMs exhibit degradation over time, indicating a lack of temporal generalization, i.e., the ability to transfer their capabilities to data beyond their training period (Biesialska et al. 2020, Lazaridou et al. 2021, Loureiro et al. 2022b). This is not surprising, as the predominant paradigm of language modeling is static (Bender et al. 2021, Lazaridou et al. 2021). (Pre-)training methods lack consideration for the temporal dimension (Biesialska et al. 2020, Dhingra et al. 2022), and evaluations commonly adhere to a temporally aligned setup, where training and test data overlap chronologically (Luu et al. 2022). In contrast, in real-life natural language processing (NLP) applications, models

[1] Code and data available at `https://github.com/IrisLuden/Beyond_Perplexity-TemporalGeneralization-DefinitionGeneration`

are often pre-trained on data from one time period and then deployed for tasks which inherently involve temporally shifted data.

To date, research on performance deterioration due to lack of temporal generalization has primarily involved analyzing the impact of factual changes in the real world on LLM performance (Lazaridou et al. 2021, Agarwal and Nenkova 2022, Dhingra et al. 2022). However, not only the facts of the world, but also the language we use to describe it changes over time. Language is a dynamic system undergoing constant evolution, marked by the emergence of new words and phrases, others falling out of use, and shifts in word meaning and usage. For instance, the interpretation of the word 'zoom' varies depending on the time and context of inquiry: before 2020, 'zoom' primarily referred to the act of making a continuous humming, buzzing, or droning sound while moving quickly. However, with the widespread use of the video conferencing platform *Zoom Video Communications*, the term gained a new meaning associated with virtual meetings.[2] Ideally, an LLM's output for, say, a text summarization task, should adapt to such a shift accordingly.

To what extent does LLM performance deteriorate over time due to semantic change, and what potential errors may arise when outdated LLMs are used for practical applications? Factual changes are easily assessed through tasks like question answering, e.g.: "Who is the president of the US?", whereas accurately evaluating an LLM's representation of word meanings demands fine grained semantic analysis. Recent studies point to a relationship between semantic change and performance deterioration as measured with (pseudo)-perplexity scores (Ishihara et al. 2022, Su et al. 2022). However, perplexity is not necessarily a proper measure of performance deterioration because a model's high perplexity on a text sequence does not necessarily imply poor performance on a downstream task (Röttger and Pierrehumbert 2021, Agarwal and Nenkova 2022). Nor is it an exhaustive measure due to its intrinsic nature: increased perplexity scores offer limited insight into the practical mistakes that an LLM can make in real-life applications. Given the potential societal impact of NLP applications, it is essential to understand how performance deterioration, particularly caused by semantic change, is manifested in LLM output.

In this work, we propose to provide deeper understanding of LLM's lexico-semantic temporal generalization through the task of *contextualized word definition generation*. The task is defined as follows. Given a context sentence c in which a target word w is used, provide the definition of the word w specifically in context c. Moving beyond perplexity, this semantic assessment offers human-interpretable insight into the ability of LLMs to infer the correct interpretation of words with changing or emerging meanings. Thus, contextualized definition generation offers a dual perspective: on the one hand, it allows for a quantitative measurement of a model's lexical understanding, and on the other hand, it gives qualitative insight into LLM's ability to process lexical semantic information through human-interpretable generated definitions. Therefore the task can be used to complement more coarse-grained measures like perplexity.

We measure temporal generalization by comparing the performance of a model on the task of contextualized word definition generation in two setups: a temporally aligned setup, where the test input overlaps in time with the model's training period, and a temporally misaligned setup, where the test input originates from a time period beyond the training period. To assess temporal generalization in the face of semantic change in particular, we differentiate between three different word categories: semantically *stable* words, semantically *changing* words, and *neologisms*. This enables the separate analysis of a model's performance for instances where semantic change is, and is not present. The LLM under scrutiny is `T5`-base, pre-trained until April 2019. `T5-base` is fine-tuned for contextualized definition generation following Huang et al. (2021). We collect a diachronic corpus $C_1 \cup C_2$ of Twitter and Reddit data, such that C_1 is temporally aligned with `T5`'s pre-training period, and C_2 is temporally shifted (i.e., it only includes data from later times periods). We apply methods for detecting lexical semantic change to this diachronic corpus to select a set of 20 stable and 20 changing target words, and use a heuristic strategy to find emerging new words. Using

[2] See `https://www.oed.com/dictionary/zoom_v2?tab=meaning_and_use#1310257030`

these three sets of target words, we create a diagnostic task of 400 <context, word> pairs in total and conduct a human evaluation study. We use human evaluation to assess whether the generated definitions for these <context, word> pairs are correct. Additionally, we analyse how `T5-base`, fine-tuned for contextualized definition generation, performs on this diagnostic task, and how this performance relates to the perplexity scores of the pre-trained model on the corresponding input.

Our results show that (i) the model's performance is adversely affected when processing temporally shifted input compared to input that is temporally aligned with the model's pre-training period, (ii) the performance deterioration is stronger for semantically changing and emerging words as opposed to semantically stable words, and (iii) cross-entropy loss and (pseudo-)perplexity scores do not reliably detect poor lexico-semantic temporal generalization. Our findings demonstrate that the proposed framework provides a promising and intuitive methodology to evaluate an LLM's ability to adapt its lexical knowledge to changing meaning conventions when semantic change has taken place. Our findings also underline the importance of assessing the capacity for temporal generalization of fine-tuned LLMs more explicitly than through perplexity scores, as perplexity is not necessarily representative of how well LLMs perform on downstream tasks.

2. Related Work

2.1 Temporal generalization

Hupkes et al. (2023) define 'good generalization' as the ability to successfully transfer representations, knowledge and strategies from past experience to new experiences. They explain that systematic generalization testing is not the status quo in the field of NLP. For decades, generalization was evaluated by training and testing models on different but similarly sampled data, assumed to be independent and identically distributed (iid). As a consequence, models do not generalize robustly in non-iid scenarios, causing performance to deteriorate when the evaluation data is different from the training data, for instance in terms of genre, topic or domain. The capacity to generalize over time is called *temporal generalization*, and is considered a form of *domain adaptation*, considering different time periods as distinct domains. Hupkes et al. (2023) propose the GenBench taxonomy describing five axes along which generalization studies can differ: the motivation, the type of generalization they aim to solve, the type of data shift they consider, the source from which this data shift originated, and the locus of the shift. We include a GenBench card for our proposed assessment in Appendix A.

LLMs are commonly trained and tested on data from overlapping time periods, while in practice, LLMs are first trained on data from one time period and thereafter applied to temporally shifted data (Sinha et al. 2021, Luu et al. 2022, Jang et al. 2022, Su et al. 2022, Hupkes et al. 2023). This phenomenon is often called *temporal misalignment.* Luu et al. (2022) showed that temporal misalignment has strong effects on performance deterioration of several LLMs for eight different tasks. They also showed that temporal (domain) adaptation by continued pre-training of the LLMs can improve performance, but that this effect is rather small compared to task specific fine-tuning on data overlapping with the test period.

Lazaridou et al. (2021) described *temporal generalization* as a model's ability to generalize well to future data from beyond their training period. 'Good generalizing' means that performance should remain consistent regardless of the time period it is tested on: if a model is capable of temporal generalization, performance should not deteriorate for data from beyond their pre-training period. To inspect LLM capacity for temporal generalization, Lazaridou et al. (2021) measure the performance deterioration of a Transformer-XL over time on temporally shifted data. Deterioration of a LLM is defined in terms of the relative performance between two setups. A "time-stratified-setup" where the LLM is tested on temporally shifted data, and a control-setup where the test data overlaps in time with the LLM's pre-training period. They compute perplexity scores for texts of different categories in both setups. The resulting 'relative perplexity' increases most for (1) texts containing emerging new words that have rarely been used in the training period, (2) texts covering politics and sports, (3) proper nouns and numbers, and (4) open-class nouns. The model's performance is also assessed for two downstream tasks: closed-book question answering, where performance is observed to decrease significantly, and reading comprehension, where performance remains consistent. This difference is somewhat surprising: Is performance deterioration for the question answering task solely caused by lack of factual information about the unseen time period? Or is it due to lack of lexico-semantic temporal generalization?

Röttger and Pierrehumbert (2021) showed that even big changes in perplexity may lead to small changes in downstream task performance. They experiment with 'temporal domain adaptation', considering different time periods as different data domains, and perform extra training of models on data from later time periods. They find that temporal adaptation improves upstream and temporal fine-tuning downstream task performance. Moreover, time-specific models generally perform better on past than on future test sets. However, they also show that adapting BERT to time does not improve performance on the downstream task over only adapting to domain. Lastly, they show that temporal adaptation captures event-driven changes in language use in the downstream task, but not necessarily those changes that are actually relevant to task performance.

Dhingra et al. (2022) observe that state-of-the-art language models are generally poor at connecting factual information to the temporal scope it applies to. LLMs are not trained to take into account the temporal context of the training data (Dhingra et al. 2022, Loureiro et al. 2022b). Understanding how facts relate to time can be seen as a prerequisite for temporal generalization: if a model generalizes its performance over time, it should be able to connect information it learns to the temporal period this information applies to. Again, this raises the question of what causes performance deterioration over time in question answering tasks: Has the model not been exposed to the correct facts during training? Or is the model architecture simply not designed to take temporality into account?

A few attempts have been made to take the temporal dimension into account when training LLMs. Rosin et al. (2022) make an interesting attempt towards training "time-aware" language models. They propose to train what they call a 'temporal contextual language model', which uses the timestamp of a text as an additional context to the texts. In this way, the model is not only trained to predict the text sequences based on the context words, but also on the time at which the text was written. They show a positive effect on performance in the "sentence time prediction task" and also on semantic change detection.

Loureiro et al. (2022b) train a set of language models called TimeLMS that are specialized on diachronic twitter data. They first pre-train a 'base' RoBERTa model on data up to 2019. Next, they continually train a new model from the base model every three months. The process of updating the base model follows the same procedure as the initial pre-training. Their work allows the NLP community to use up-to-date LMs of any period of time, which can be useful to compare performance in the quest to alternatives to the current static language modeling paradigm.

Jang et al. (2022) introduce a benchmark called TemporalWiki that is collected by an algorithm that detects which facts have been newly added to Wikipedia at a certain point in time. This benchmark is used to continually train LMs only on an 'updated' portion of English Wikipedia data, such that a LM only needs to be 'updated' on a smaller portion of data that is considered relevant because it contains new information. This would reduce the amount of extra training of an outdated LM. They find that training an LM on this TemporalWiki data set achieves better perplexity than on the entire Wikipedia with 12 times less computational costs.

2.2 Temporal generalization and semantic change

Several works have indicated that semantic change is related to temporal performance deterioration. Su et al. (2022) examine the impact of semantically changing words on the performance of a language model. They do so by showing that extra training on data containing semantically changed words, opposed to just a random set of words, improves the perplexity scores of pre-trained language models significantly. Their method even yields performance improvement over domain adaptation methods on two different pre-trained language models and four data sets. This indicates that language models suffer from performance deterioration due to lack of understanding of semantically shifted words.

Ishihara et al. (2022) show a negative correlation between semantic change and perplexity for Word2Vec and RoBERTa. They show that a large time-series performance degradation occurs in the years when the so-called *semantic shift stability* is smaller. The degree of semantic shift is approximated by performing lexical semantic change detection between Word2Vec models created from corpora of different time periods. A low degree of semantic shift between two time periods implies semantic shift stability between these time periods.

Loureiro et al. (2022a) present a benchmark called TempoWic. This benchmark consists of tweets containing 'trending words', as trendiness of a word is an indicator of semantic change. For each trending word, the benchmark provides ground truths about whether its meaning is identical or different in two tweets. This benchmark can thus be used to assess how well a model predicts whether the meaning of a target word is identical, or different in two different contexts.

Eisenschlos et al. (2023) experiment by 'mimicking' the scenario that new words enter a language. They test how the performance of an LLM is impacted on the task of part-of-speech tagging, when the LLM is not acquainted with the key verb phrase in a sequence to be tagged. They simulate neologisms by replacing existing nouns in the agent/object relation with a verb with made-up words. They show that the performance of the model is decreasingly worse when it is presented with input containing unknown words. However, performance increases tremendously if the LLMs are provided a definition for the new words, indicating that LLMs are capable of temporally generalizing as long as they are provided with the necessary information.

In sum, prior work has indicated that LLMs suffer from temporal performance deterioration, and that this deterioration is at least partially a consequence of temporal semantic change. So far, results are mainly based on perplexity scores, bet these do not give insight into the possible semantic mistakes a LLM can make when it is outdated. A lexico-semantic analysis as to why an LLM deteriorates over time is lacking.

3. Lexico-Semantic Temporal Generalization: Hypotheses

We propose to assess the temporal generalization of a language model in the face of lexical semantic change by exploiting the task of contextualized word definition generation. Recall that the task, exemplified in Figure 1, can be defined as follows: given a target word w in a context c, the task consists in generating a dictionary-style natural language definition of w's meaning in c. Autoregressive LLMs can be directed to perform this task through methods such as fine-tuning or few-shot in-context learning, and their performance can be evaluated using reference-based natural language generation (NLG) measures (when reference contextualized definitions exist) or human evaluation. For instance, Mickus et al. (2019) fine-tune a Transformer model (Vaswani et al. 2017) for this task, and use perplexity to evaluate the model.

More recently, Giulianelli et al. (2023) use `Flan-T5` to generate word definitions using prompts consisting of an example usage followed by an instruction, and they evaluate their approach with human evaluation.

Definition generation models have proved promising, but they may still struggle with part-of-speech mismatching, lack of fluency, finding the appropriate specificity for a definition, and avoiding self-referential definitions (Noraset et al. 2017) and hallucinations (Mickus et al. 2022). Yet, we argue that the in-depth semantic assessment offered by this task provides interpretable insights into the possible mistakes a model may perpetrate due to meaning shift, and therefore is a useful complement to more coarse-grained measures like perplexity scores.

Given a model trained on data up to time t, we assess its temporal generalization by testing its performance on the task of contextualized word definition generation using test sentences originating after its training period. In particular, we compare its performance in a standard *temporally aligned*

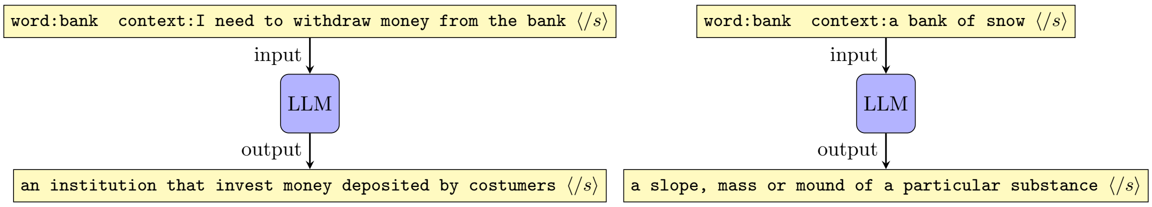

word:bank context:I need to withdraw money from the bank ⟨/s⟩
input
LLM
output
an institution that invest money deposited by costumers ⟨/s⟩
word:bank context:a bank of snow ⟨/s⟩
input
LLM
output
a slope, mass or mound of a particular substance ⟨/s⟩

Figure 1: Two examples of contextualized word definition generation.

setup, where all datapoints in both train and test sets have a timestamp $t^{-} <= t$, to its performance in a *temporally misaligned* setup, where all the sentences in the test set originate at some time $t^{+} > t$. A model is able to temporally generalise in the presence of semantic shift if its performance does not deteriorate in the temporally misaligned setup compared to the standard temporally aligned scenario.

As mentioned in the Introduction, we distinguish three categories of target words for the task of contextualized word definition generation: Semantically *stable* words are words that have not undergone semantic change since time t; semantically *changing* words are words that have undergone semantic change since time t; finally, *emerging* words are words that have either newly entered a language's vocabulary after t or occur significantly more frequently since t. These three categories allow us to formulate three hypotheses:

Hypothesis 1 *Performance on the contextualized word definition generation task will be higher for texts with timestamps $t^{-} <= t$ than for texts with timestamps $t^{+} > t$ for any of the target word categories.* That is, we hypothesise that models will show poor temporal generalization: a model's understanding of word meaning will decline when word usages are sampled from temporally shifted time periods.

Hypothesis 2 *Performance deterioration in the temporally misaligned setup will be more prevalent for changing target words than for stable target words.* Since previous studies have indicated that there is a correlation between semantic change and performance deterioration, we expect that semantic change will impact a model's performance on contextualized word definition generation.

Hypothesis 3 *Performance on emerging target words will be lower than for stable target words.* Since emerging words are words that have newly appeared or rapidly increased in frequency since t, it is likely that the model has not been exposed to many training instances containing the emerging target word, making it more difficult to generate adequate definitions for them.

Besides quantitative scores, the task also enables qualitative analysis of the generated content: the generated definitions themselves reflect "semantic information" that a model emulates. Analysing output in the form of definitions makes it possible to get human-interpretable insight into the implicit semantic information that a model represents of the words under investigation. This can also provide insight into the possible subtle semantic mistakes that a model can make when applied to other (generative) tasks such as text summarization.

4. Experimental Setup

In this section, we describe the setup in which we explore the use of the task of contextualized word definition generation to assess lexico-semantic temporal generalization. First, we collect two corpora: one temporally aligned with the model's training period and another temporally misaligned. Using these corpora, we gather target words for each of the three categories (stable, changing, and emerging) with the assistance of a lexical semantic change detection system. For each target word, we also extract a selection of usages to obtain ⟨word, context⟩ pairs that can be used as input for the task of contextualized word definition generation. Next, we detail the pre-trained LLM and the fine-tuning method employed to perform the task. Finally, we explain the evaluation process. We will make available both the corpora we created and our experimental code upon publication.

4.1 Corpus creation

We create two corpora, C_1 and C_2, to identify candidate target words for our three categories of interest: stable, changing, and emerging words. As explained in more detailed in Section 4.3, we experiment with `T5-base` (Raffel et al. 2020). Since `T5-base` was pre-trained on data until April 2019, we set t as the 1st of May 2019. We set t_{end} at the latest possible time of conducting the experiments, February 2023. Thus, our temporally shifted corpus, C_2, covers data from 46 months in total. To keep the periods of both corpora equal, we set $t_{start} :=$ 1 July 2015, such that C_1 also covers data over 46 months.

The corpora are constructed from Tweets and Reddit posts and comments. We use the pipeline developed by Loureiro et al. (2022b) to request Tweets using the academic Twitter API. They have published a pipeline that allows users to request tweets per month, filter out tweets by unauthorized users, and anonymify the user accounts. This yielded a total of 4,5 million tweets. We use the Pushshift Reddit API (Baumgartner et al. 2020) to collect sentences from Reddit. For each day between July 2015 to February 2023, we request at most 500 posts and 500 comments. Only the posts and comments that consist of at least 10 words (and contain at least one English stop word from `nltk.stopwords`, following Loureiro et al. (2022b)) are included. This results in roughly one million posts and comments. We make sure that C_1 and C_2 are (roughly) the same size.

Data pre-processing and cleaning We use NLTK (Loper and Bird 2002) to split documents into sentences and to tokenize sentences into words.[3] Words are stripped from punctuation and made lower case, and emoji's are removed. Basic statistics for the two corpora are provided in Table 1.

dataset		start date	end date	# docs	# words
C_1		01.07.2015	01.04.2019	3.4M	70M
	Twitter			2.2M	46.2M
	Reddit			1.2M	23.9M
C_2		01.05.2019	01.02.2023	3.4 M	79M
	Twitter			2.2M	55.6M
	Reddit			1.2M	23.6M

Table 1: Corpus statistics.

4.2 Target word selection

We design the task for the three word categories: *stable*, *changing*, and *emerging* target words.

We consider as candidate stable and changing target words any words in the shared vocabulary of C_1 and C_2 which have an entry in the WordNet database (Fellbaum 1998),[4] do not contain any digits (e.g., a term like '2022' is excluded), and are not proper nouns or abbreviations. We then apply one of the best-performing lexical semantic change detection systems according to Schlechtweg et al. (2019): `SGNS+OP+CD`. This system uses SkipGram with Negative Sampling (SGNS) to construct a vector space model for each corpus separately. The two vector space models are aligned using Orthogonal Procrustes (OP). For each word, the semantic change score is calculated by computing the cosine distance (CD) between its word embedding from each of the aligned vector space models. This results in a semantic change score per word.

[3] In particular, we use NLTK's `sent_tokenize`, `TreebankWordTokenizer` and `nltk.TweetTokenizer`.

[4] Following Su et al. (2022).

Stable target words To collect a set of *stable target words*, we use a random selection of 20 words that have a semantic change score below 0.25, displayed in Table 2a. Examples of stable target words are LOOK, SETTINGS and IDEA.

	Word	CD
1	look	0.12
2	lose	0.12
3	player	0.13
4	morning	0.14
5	population	0.17
6	option	0.17
7	idea	0.17
8	settings	0.18
9	opinions	0.18
10	statement	0.19
11	families	0.20
12	realise	0.20
13	community	0.22
14	asparagus	0.22
15	art	0.22
16	talks	0.22
17	beginning	0.22
18	outcome	0.22
19	groceries	0.22
20	performance	0.22

(a) Stable

	Target Word	CD
1	corona	0.98
2	lockdown	0.96
3	manifesting	0.92
4	closeness	0.91
5	pandemic	0.90
6	quarantine	0.88
7	navigator	0.86
8	distancing	0.83
9	ape	0.81
10	checkmate	0.79
11	masking	0.78
12	peacock	0.78
13	polygon	0.76
14	anchor	0.75
15	shanks	0.74
16	tracing	0.73
17	pinks	0.72
18	moot	0.72
19	hag	0.72
20	yacht	0.72

(b) Changing

	Emerging word
1	copium
2	covidiots
3	plandemic
4	vaxed
5	gatekeeping
6	grifting
7	gaslight
8	non-binary
9	femboy
10	quarantining
11	covid
12	transphobe
13	simp
14	wokeness
15	sapphic
16	spreader
17	goated
18	k-pop
19	vax
20	anti-vax

(c) Emerging

Table 2: Target words for each category with their semantic change score.

Changing target words To collect a set of changing target words, we first make a pre-selection according to the *trending scores* of this vocabulary, as word trendiness is an indicator of semantic change (Chen et al. 2021, Loureiro et al. 2022a)). The trending score is defined as follows:

$$score(w) = \frac{f_{w,C_2} - f_{w,C_1}}{f_{w,C_2} + k} \tag{1}$$

where f_{w,C_i} is the highest monthly frequency of word w in corpus C_i. k is a normalization term used to mitigate the frequency of highly-frequent terms in the recent data sets. We compute the semantic change score for all words with a trending score above 1. Of these, the top 20 words with the highest change score are selected as *semantically changing* target words. Table 2 displays the 20 selected target words of each category and their semantic change scores.

Many of the *changing target words* were related to the COVID-19 outbreak, e.g., CORONA, LOCKDOWN, PANDEMIC, QUARANTINE and DISTANCING. This is not surprising, as the COVID-19 outbreak happened by the end of 2019, which started after the pre-training period of `T5-base`. For instance, before the outbreak, *corona* was either used to refer to a Mexican beer brand, or to a city in the US. The word MANIFESTING has likely received a high semantic change score because of the emergence of a new sense: a definition was added to the Urban Dictionary on December 6th of 2020[5], defining it as 'a term used by subliminal users meaning to hope for a desire until it comes true using

[5] See `https://www.urbandictionary.com/define.php?term=manifesting`

the law of attraction'. In C_1, manifesting was probably used as simply the present participle of the verb To manifest, which describes the process of making something visible or apparent.[6]

Other changing target words likely owe their high change score to a growing prevalence of the word due to cultural events. This is particularly prominent for words that are used as proper names, but were not detected by our proper-name filter because they can also be common nouns. For example: Polygon is increasingly used to refer to an online gaming platform; shanks is increasingly used to refer to a Japanese Manga character in C_2; hag is frequently used to refer to the Dutch football coach *Eric Ten Hag*: 40% of the sentences in C_2 contain the n-gram Ten Hag, compared to less than 1% in C_1.

Emerging target words We define *emerging words* as words that either (i) have a document frequency in C_2 of at least 50, while having a document frequency of 0 in C_1, or (ii) have a document frequency that is at least five times as much in C_2 compared to C_1. This resulted in a total of 1585 emerging words. Since newly emerging words are likely not present in the WordNet database, we manually select 20 words that (i) do not contain any digits, (ii) are not named entities (i.e. places, persons, brands) (iii) are not abbreviations.

Many of the emerging target words, like the changing target words, relate to COVID-19: covid, covidiots, plandemic, vaxed, covid, spreader, vax, anti-vax. Other emerging words relate to gender identity: non-binary, femboy, sapphic, transphobe. A particularly interesting emerging word is goated, which is an example of grammatization from the noun goat, which was initially an abbreviation for Greatest Of All Time. Emerging words that originate from blends are copium (cope + opium), covidiots (covid + idiot), and plandemic (plan + pandemic). The word gaslight is the result of the concatenation of two already existing words.[7]

Diagnostic dataset For each target word, and each corpus in which they occur, 4 sentences are randomly sampled as input sentences to the task.[8] This results in a total of 160 instances of stable target words, 160 instances of changing target words, and 80 instances of emerging words (only from C_2). This provides us with a total of 400 context-target pairs that can be used as input to the task. To illustrate, we list six examples, two of each category, in Table 3.

word	**context**
population	they're experimenting on the population and it needs to stop.
population	approximately 16.5 million tourists visit greece each year. that's more than the entire population of greece!
distancing	i can feel you distancing from me and it sucks because i only got eyes for you
distancing	social distancing and covid 19 health precautions may be hard to abide by when the epl resumes next week
wokeness	also being able to read and write well and having some reasonable sense of history would be powerful inoculation against wokeness.
wokeness	an actual good tweet against the wokeness only gets 2.5k likes but your reply guy ugly tweets against republicans get thousands and thousands of likes?

Table 3: Target words and example context sentences.

4.3 Pre-trained model and fine-tuning

In our setup, the model under investigation is `T5-base` (Raffel et al. 2020). There are three main practical reasons to choose this model. Since `T5` is a sequence-to-sequence model, it is compatible

[6] See `https://www.oxfordlearnersdictionaries.com/definition/english/manifest_1?q=manifesting`

[7] Other notable examples not selected as target words, include the acronym prod for product, and the abbreviation ima for I'm going to.

[8] We acknowledge that random selection may not be optimal. Section 7 includes a discussion of this point.

with the contextualized definition generation task. Secondly, the time period from which its pre-training data originates is well documented (opposed to some other large pre-trained models where the pre-training data is undisclosed). Third, since the model was pre-trained on data until 2019, there exists enough temporally shifted data on which the model can be tested on temporal generalization.

Generating definition with a suitable level of specificity can be tricky (Noraset et al. 2017, Mickus et al. 2019). In our setup, we choose to fine-tune `T5-base` for contextualized definition generation according to the method proposed by Huang et al. (2021), who designed a definition modeling procedure optimized to produce definitions of appropriate specificity. Huang et al. (2021) fine-tune three `T5` models—`T5-base`, `T5-specific`, and `T5-general`—which are combined to select the definition with the most appropriate level of specificity given the provided context. `T5-base` is the main language model under analysis, a version of `T5` fine-tuned to generate definitions for a given target word and context. `T5-specific` is used as an over-specificity estimator. It is fine-tuned to generate a usage example for a given target word $w*$ conditioned on a reference definition. Under this model, pairs of highly specific definitions and respective contexts are assigned high probability. The third model, `T5-general`, is used as an under-specificity estimator. It is fine-tuned to generate a definition conditioned on a target without any usage examples. Under this model, pairs of generic definitions and the respective target words are assigned a high probability.

The three models are fine-tuned to minimize the cross-entropy loss for their respective tasks on the Oxford dataset, consisting of definitions and usage examples collected by Gadetsky et al. (2018) from the Oxford Dictionary. The target words that also occur in the Oxford training data set were removed. For specificity-enhanced generation, n definitions are decoded from the main model, `T5-base`, and then re-ranked according to a linear combination of the probability scores assigned by the three models. The resulting definition generator will henceforth be referred to as `T5-base-DG`.

4.4 Evaluation

We evaluate the quality of the definitions generated by `T5-base-DG` given a word-context pair with help of human annotation. This allows us to obtain accuracy scores for each target word category (stable, changing, and emerging words), and each time period. Furthermore, we investigate how the quality of the generated definitions on a word-context input relates to the (pseudo)perplexity and log-likelihoods on these input context sentences.

Human evaluation To determine whether the generated definitions are correct, we conduct human evaluation on the 400 generated definitions for a given target word and usage example. Three human annotators (fluent English speakers) were presented with a total of 400 (word, example sentence, definition) triplets. For each triplet, the annotators judged correctness of the generated definition on a four-point scale between 0 and 3, where a score of **3** corresponds to a completely correct definition, i.e., one that is both truthful and fluent, and a score of **0** corresponds to a completely incorrect definition. A special case for incorrect definitions is self-reference: a definition is self-referring whenever it includes the target word itself to define the target word. In this case, the annotators were instructed to assign the score **-10**. The judgements were aggregated via majority vote into binary correctness scores, where the labels **-10, 0, 1** are considered *incorrect*, and the labels **2, 3** are considered *correct*. This allows computing quality in terms of the percentage of correct definitions. Full annotation guidelines can be found in the Appendix 7.3. The inter-rater agreement is measured using Krippendorff's α coefficient, which quantifies the extent to which the observed agreement goes beyond what would be expected by chance. The coefficient ranges from 0 to 1, with higher values indicating greater agreement.

Perplexity We also investigate the relation between model perplexity of usage examples and definition quality. We are interested to see whether high model perplexity on sentences containing a target word indicates performance deterioration when these sentences are used as usage examples in the definition generation task. Besides perplexity, we also compute the cross-entropy loss, and pseudo-log-likelihood of the model on the context sentence. Perplexity is a measure of how well a language model predicts sequences of words, with lower values indicating better performance. Perplexity is derived from the cross-entropy loss, by taking the exponentiation of the cross-entropy loss, and yields a value that is easier to comprehend and compare.

Since an LLM can be considered a probability distribution over all possible text sequences of a language, the cross entropy loss ($\mathtt{H}$) and perplexity ($\mathtt{PPL}$) and can be used to estimate how well the language model predicts a sequence of words S (Ranjan et al. 2016). The cross entropy measures the degree of 'uncertainty' when encountering a text sequence, while the perplexity measures the degree of 'surprisal' a model has in predicting a text. The two measures are closely related, as $\mathtt{PPL}(S) = e^{\mathtt{H}(S)}$.

Let $\mathbb{P}(S)$ denote a language model's probability of the sequence $S = w_1, w_2, \ldots, w_n$, where each w_i is a word in the vocabulary, and $n \in \mathbb{N}$ is the number of words in the sequence. The higher the cross entropy loss, the more surprised the model is to encounter the sequence S. The cross-entropy score is calculated by: $\mathtt{H}(S) = -\frac{1}{n} \log \mathbb{P}(S)$. Perplexity is defined as the language model's inverse probability of the sequence S, normalized by n. The $\mathtt{PPL}$ score of a text $S = w_1, w_2, \ldots, w_n$ is calculated using $\mathtt{PPL}(S) := \sqrt[n]{\frac{1}{\mathbb{P}(S)}} \equiv \frac{1}{n} \log \mathbb{P}(S)$. The higher the perplexity score for a given sequence is, the more 'surprised' the model is to encounter this sequence.

The perplexity and cross-entropy loss scores of a text sequence disregard how the presence of each word in the sequence contributes to the sequence likelihoods. For instance, it could be the case that only one word in the sequence is particularly unlikely, while the rest of the sequence is relatively likely. Therefore we also compute the pseudo-log-likelihood, which subsequently computes the cross-entropy loss of each individual term within the entire sequence (Salazar et al. 2020). A model's pseudo-perplexity for a document T containing N tokens is: $\mathtt{PPPL}(T) := \exp -\frac{1}{N} \sum_{S \in T} \mathtt{PLL(S)}$ where $\mathtt{PLL}$ is the *pseudo-log-likelihood* of the sequence $S = (w_1, \ldots w_n)$: $\mathtt{PLL}(S) = \sum_{t=1}^{n} \log \mathbb{P}(w_t | S_{\backslash w_t})$, where $S_{\backslash w_t}$ represents the sentence where w_t is masked.

Additionally, we compute the cross-entropy loss for the masked-word-prediction task to get an indication how the appearance of a target word w_t in the context sentence S contributes to the sentence perplexity: $\textsc{Loss}(w_t, S) := -\log \mathbb{P}(w_t | S_{\backslash w_t})$. Again, this is computed by replacing the target word in the sentence with the special `<extra_id_1>` mask token (which results in the masked sequence $S_{\backslash w_t}$), and computing the model's cross-entropy loss for predicting the target word in that position. We compare these scores along three dimensions: (1) human-annotated definition quality, (2) the time period of the usage example and (3) the target word category.

5. Results

Using the definition quality scores obtained with help of human evaluation, we calculate the performance deterioration of `T5-base-DG` for each category of target words. Comparing definition quality for the stable vs. changing and emerging target words provides insight into how word usage change impacts performance deterioration. Furthermore, we compare the performance on the contextualized definition generation task with the cross-entropy loss and perplexity scores of `T5-base` for the same example contexts to investigate the relation between these standard metrics of model fit and lexico-semantic generalization.

5.1 Performance deterioration on the contextualized definition generation task

Human judgements of definition quality per category, aggregated by majority vote, are shown in Table 4. Krippendorff's α inter-rater agreement is 0.62.[9] The performance of `T5-base-DG` deteriorates by 19% (from 52.5% to 42.5%) between C_1 and C_2, confirming our first hypothesis that the performance on the input $\langle word, context \rangle$ pairs from C_1 is higher than for those from C_2. Performance

[9] And 0.68 if we reduce the labels to four, mapping the -10 judgement to 0 (incorrect).

deterioration is especially strong for semantically changed words, with a decrease in performance of 36.7% (from 37.5% to 23.75%), compared to 7.5% (from 66.25% to 61.25%) for the stable target words. This confirms our second hypothesis that the performance deterioration is more prevalent for semantically changing words compared to stable words. Finally, as expected, we observe a drastic drop in generation quality for emerging words, compared to both stable and changing words, confirming our third hypothesis that the performance deterioration is stronger for emerging words compared to stable words.[10]

Overall, performance deterioration between C_1 and C_2 in all categories reveals that `T5-base-DG` lacks strong temporal generalization capabilities.

Beyond these main trends, two other aspects of the human evaluation results are worth a note. First, while performance deterioration is stronger for changing and emerging words, it is still visible for stable words. This implies that lexico-semantic temporal generalization can suffer not only from clear-cut semantic shifts occurring for a certain target word, but also when context words in the usage examples change their meaning (which we are currently not tracking in our setup) or when the word usage distributions change in more subtle, less semantically determined ways over time. Second, in the time period C_1, which is temporally aligned with the training data by design, we find generation quality to be lower for changing words than for stable words. This suggests the lexical meaning of the changing words under analysis may be inherently more difficult for the model to grasp due to other properties, such as their degree of specificity. Nevertheless, the drastic decrease in performance for the emerging and changing words, substantially stronger than that of stable words, indicates that semantic change and performance deterioration are related.

Category	C_1	C_2	$C_1 \cup C_2$
stable	**66.25%**	**61.25%**	**63.75%**
changing	**37.5%**	23.75%	30.625%
stable + changing	**52.5%**	42.5%	47.5%
emerging	-	8.75%	8.75%
total	**52.5%**	31.25%	-

Table 4: Accuracy on the contextualized definition generation task, expressed as the percentage of definitions judged by human annotators as correct.

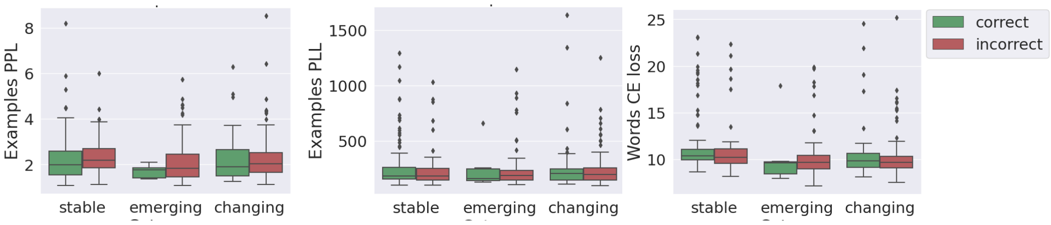
5.2 Perplexity as an indicator of temporal generalization

If perplexity were a reliable indicator of performance deterioration, we would expect the perplexity of the input sentences from C_2 to be on average higher, as these originate from a time period that the model was not trained on. However, we observe that on average the input from C_2 does not yield higher perplexity scores. We also fail to find substantial differences in perplexity between target word categories (Table 5), with stable words obtaining, in fact, slightly higher perplexity than changing and emerging ones.

If perplexity were a reliable indicator of performance deterioration, we would also expect usage examples with higher perplexity scores to correspond to incorrect definitions. Instead, the correlation between the perplexity scores on usage examples and the correctness of the respective definitions is either not significant or very moderate (Table 6). Figure 2 illustrates how the perplexities are distributed over usage examples for each word category, split by correctness label. We can indeed see that some input sentences have relatively high perplexity but still yielded correct definitions, and vice versa.

Overall, perplexity seems to be a poor indicator of lexico-semantic temporal generalization.

[10] These trends, that performance deterioration is stronger for changing and emerging target words than for stable target words, persist when aggregating the judgements by consensus voting. The results aggregated by consensus voting can be found in Appendix D



(a) Sentence `PPL` (b) Sentence `PLL` (c) Word prediction `CEL`

Figure 2: Intrinsic scores of the perplexity (PLL), pseudo-log-likelihood (PLL), and target word prediction cross-entropy loss (CEL), versus the correctness of the generated definition

6. Qualitative Analysis

We conduct a detailed analysis of the generated definitions of `T5-base-DG`. First, we illustrate that the model `T5-base-DG` can generate context-specific definitions for input $\langle word, context \rangle$ pairs. Next, we discuss the errors that occur for the changing and emerging target words. This qualitative error analysis shows us what the possible reasons are for incorrectly generated definitions.

Success analysis: context-specific definitions We begin with two examples which demonstrate that `T5-base-DG` is capable of generating context-specific definitions, displayed in Table 7. In the first usage example of the stable word LOOK, example *a*), the word is used in the context of LOOK OUT TO SEE. . . , for which the generated definition is 'expect something to happen or be the case'. A different use of the word LOOK, as in 'LOOKS LIKE. . . ', is displayed in example *b*); for this usage type, the generated definition is 'have the appearance of being'. Both definitions are judged to be correct by all three annotators. Similarly, example *c*) of the stable word OPTION shows that `T5-base-DG` has generated a remarkably precise, fluent and correct definition, namely 'an item in a list or other collection of data displayed on the screen which can be selected'. For example *d*) of OPTION, the definition 'an idea or way of doing something' is not judged sufficiently correct by the annotators. Nonetheless, even though the generated definition is not sufficiently precise, we can still distinguish a different word sense of OPTION compared to example *c*). Examples *e*) and *f*) show correctly generated definitions for the changing target word LOCKDOWN. Definition *e*), 'a state of isolation or seclusion by the police', refers to a situation where someone is under strict control or confinement, potentially as a response to a legal or security issue. Definition *f*), 'a period of isolation or separation from other people' for the input from C_2 does not necessarily imply police enforcement but rather emphasizes the idea of being confined or restricted, leading to the person's son feeling bored and lonely. Thus, these cases exemplify that model is capable of generating context-specific definitions for a target word.

Error analysis: changing target words Table 8 displays examples of two changing target words (CORONA and POLYGON) and their generated definitions. These target words are examples of words

Category	Corpus	Sentence `CE loss`	Sentence `PPL`	Word prediction `CE loss`
stable	C_1	**0.757**	**2.399**	11.34
	C_2	0.734	2.303	**11.69**
changing	C_1	0.695	2.177	10.69
	C_2	0.700	2.190	11.40
emerging	C_2	0.688	2.184	10.82

Table 5: Average scores of cross-entropy and perplexity on the usages per category.

	Point Biserial	Spearman
Word prediction `CE` loss	0.13 (0.009)	0.14 (0.000)
Sentences `PPL`	-0.02 (0.752)	-0.04 (0.336)
Sentences `CE` loss	-0.03 (0.551)	-0.04 (0.336)
Sentences `PLL`	0.08 (0.117)	0.03 (0.521)

Table 6: Correlations with correctness of the contextualized definition generation task.

	Target	LOOK
a)	**Example C_1**	in heroes and generals there you have limited sight as gunner but you can *look* out to see more with the danger of getting shot.
	Definition	expect something to happen or be the case
	Judgement	2, 2, 3
b)	**Example C_2**	*looks* like russia had exposed to the world that they are not a military might.
	Definition	have the appearance of being
	Judgement	3, 3, 3
	Target	OPTION
c)	**Example C_2**	selecting the aa *option* is now in loadout screen as opposed to being a modification.
	Definition	an item in a list or other collection of data displayed on the screen which can be selected
	Judgement	3, 3, 3
d)	**Example C_2**	also if an address is required then is there another *option* other than providing my address?
	Definition	an idea or way of doing something
	Judgement	1, 1, 1
	Target	LOCKDOWN
e)	**Example C_1**	imagine being on *lockdown* up until you get to college ... boom !
	Definition	a state of isolation or seclusion by the police
	Judgement	3, 3, 2
f)	**Example C_2**	this *lockdown* needs to end foreal my son is so bored and lonely now he is demanding for a baby sister ...
	Definition	a period of isolation or separation from other people
	Judgement	3, 2, 3

Table 7: Examples of context-specific generated definitions, with human evaluation judgements.

for which a new meaning has emerged in $t^+ > t$. One would expect that the generated definitions would be correct for the $\langle word, context \rangle$ pairs retrieved from C_1, as these sentences originate from a corpus that is temporally aligned the model's pre-training period. However, the generated definitions are by and large incorrect for both C_1 and C_2. We choose to discuss these two examples, because they clearly illustrate three different phenomena.

The first phenomenon is that we observe that the poor quality of the generated definitions from C_1 input seem to be of a different nature than the definitions generated from C_2 input. The definitions for $\langle word, context \rangle$ pairs from C_1 do display knowledge of semantic relatedness to the intended sense of the target word (even though the generated definitions incorrect). For instance, for the word CORONA, three of the input usages refer to the Mexican beer 'Corona', and one usage refers to the city of Corona ('I live in corona. . . '). Thus, in each of these examples, the generated definition is semantically related to the correct sense. Likewise, generated definitions from the usages of C_1 for POLYGON are each semantically related to the correct sense of the target word referring to 'a plane figure with many straight sides and the same number of angles'.[11] Even though the generated definitions for the usages from C_1 are factually incorrect, they are still semantically related to the correct sense of the target word.

The second phenomenon that we observe is that when generating a definition for a changing target word with a new sense, the model relies largely on the provided context words. This is for instance apparent in 'a strain of arnovirusses . . . ' to define CORONA, and a 'computer graphic . . . ' to define POLYGON. It seems that here, `T5-base-DG` does detect a 'new' sense. However, the model seems to rely too much on the given contexts words to generate the definition, resulting in the generation of an incorrect definition.

The third phenomenon that we observe is that the definitions generated for the $\langle word, context \rangle$ pairs from C_2, remain semantically related to the pre-existing sense of the target word from $t^- < t$. For instance, the definition 'a cigar' is likely generated because CORONA is also a brand of Havana cigar.[12] Thus, the definition refers to a different, pre-existing sense of the target word. Likewise, in the case of POLYGON, two of the definitions generated for usages from C_2 are still semantically related to the pre-existing sense of the target word. These definitions are: 'more than three dimensional elements' and 'many-dimensional'. These definitions once again refer to a geometrical plane figure, while the correct definition should refer to the newer sense of POLYGON, which refers to an entertainment website. It seems that for these examples, the model does not recognise a new word sense, but rather keeps relying on prior knowledge about the target word to generate a definition.

To summarize, the incorrect definitions generated for input from C_1 display similarity to the correct word sense (phenomenon 1), while the incorrect definitions generated on input from C_2 display similarity to the context words (phenomenon 2), or display similarity to the pre-existing sense of the target word (phenomenon 3).

Success analysis: emerging target words Out of the 80 usages of emerging words, only 7 of the generated definitions were judged to be correct.

Let us first discuss why these 7 instances were actually correct. Two of these are for usages of GASLIGHT, and two of ANTI-VAX. The two correctly generated definitions for ANTI-VAX are (1) 'antipathy or aversion to vax', and (2) 'a person who has no vaccinations or is actively anti-viral'. The first definition is correct, as it is fluent and factual. However, the model still incorrectly defines VAX in all cases, with definitions like 'a disease caused by an infection of the vagina', and 'ask for or obtain as a vaex'. In contrast, definition (2) is surprisingly correct, apart from the fact that the term 'anti-viral' is slightly ambiguous. The correctness of this definition can be explained by the informativeness of the example sentence that was provided, which was: 'swagenknecht okay go ahead you call this guy anti-vax because he is not vaccinated!'. This example sentence is largely a definitial sentence itself, as it explicitly states why a person is anti-vax. Incorrectly generated for ANTI-VAX are (3) 'exaggerated or anti-vox', and (4) 'hostile or obnoxious'. Thus, `T5-base-DG` is capable of generating a correct definition for an emerging words if the provided usage provides with sufficient information to deduce the defininition.

The correct definitions for GASLIGHT were (1) 'the light of a gaslamp', and (2) 'manipulate (someone) by psychological means into doubting their feelings'. Both definitions refer to a different sense of the word GASLIGHT; the first being the traditional use of GAS + LIGHT, while the latter corresponds to the emerged sense, which is defined correctly by the generated definition. Contrary to ANTI-VAX, the example sentences of GASLIGHT are not as informative that the model can copy

[11] See `https://www.oed.com/dictionary/polygon_n?tab=factsheet#29555477`

[12] See: `https://www.britannica.com/topic/corona-cigar`

	corona C_1	Correct?
1.	a cocktail made with aromatic spices and fruit juice	no
2.	a deep red or yellowish-brown colour	no
3.	a cold drink served with drinks such as fruit or vegetables	no
4.	a small lake or valley	no
	corona C_2	
1.	the identification of a kite or other mammal by its markings and colours	no
2.	a cigar	no
3.	a divinely conferred blessing or beneficence	no
4.	a strain of arnoviruses found in many tropical and subtropical areas	no
	polygon C_1	
1.	a solid or cylindrical object having at least three straight sides and angles	yes
2.	more than three dimensional parts or elements	no
3.	a very large number or amount	no
4.	a word or phrase used by several people	no
	polygon C_2	
1.	a three-dimensional recreation in which players use two or more lines to move around one another	no
2.	many-dimensional	no
3.	denoting a conceptual system in which data is represented by two or more discrete units	no
4.	a computer graphic or display device that supports several different configurations	no

Table 8: Changing target words and their generated definitions

the definition from the example sentence. An explanation for this is that the term gaslight, and its corresponding emerging sense of 'manipulate (someone) by psychological means into doubting their feelings' is not completely new, as it originates from the British theater play 'Gas Light' of 1938, and was added to the Urban Dictionary in 2009.[13] This makes it likely that this sense of gaslight was already used in the pre-training corpus of `T5-base`. Thus, the generated definitions for emerging word gaslight are likely correct because these usages do not actually display new language use.

Error analysis: emerging target words Examples of incorrectly generated definitions can be viewed in table 9 below. These illustrate some other common errors that `T5-base-DG` makes.

Firstly, when presented with new words, `T5-base-DG` in turn also produces definitions containing non-existing new words. This was the case for vaex when defining vax, for anti-vox when definition anti-vax, plandelia and plandisone when defining plandemic, and a-femboy to define femboy.

Secondly, many of the emerging words trigger some weak or strong form of self reference. This seems to happen more than for the stable target words. This was the case for gatekeeping, spreader, plandemic, non-binary, femboy, wokeness and quarantining.

Thirdly, some of the incorrect generated definitions reflect an implicit polarity (positivity or negativity) towards the target word. This polarity seems to be inferred from the provided context (the example sentence). For instance, the generated definitions for the word simp are considerably negative: 'a weak or ineffectual person', 'a stupid or contemptible person', 'a servile or impudent woman', and 'an impudent or insincere man'. In fact, according to the online dictionary, the definition of simp is: 'a slang insult for men who are seen as too attentive and submissive to women,

[13] See: `https://www.urbandictionary.com/define.php?term=Gaslighting`; `https://www.washingtonpost.com/wellness/2022/04/15/gaslighting-definition-relationship-abuse-response/`

especially out of a failed hope of winning some entitled sexual attention or activity from them'.[14] Likewise, FEMBOY is defined as 'a lame or mischievous person', while in fact, it means 'a young, usually cisgender male who displays traditionally feminine characteristics'.[15] Thus, the model does catch on to the negative polarity of the context, however, attributes incorrect qualities to the word. Likewise, the definitions for COVID reflect negative ('a term of abuse') or positive ('a term of endearment') connotations, depending on the sentiment of the input context. Arguably, these definitions display a form of bias.

	COVID
1.	used as a general term of abuse
2.	divergence from sex in the sexual activity of women
3.	used as a term of endearment
4.	an entertaining or amusing person
	K-POP
1.	denoting a category of words in radio and television programmes that are intended to attract attention
2.	pop music or dance to a popular song of australian origin
3.	a style of popular music intended for people who are secretly seeking to attract attention
4.	relating to or denoting unrestrained folk music of us black origin
	FEMBOY
1.	a showy or frivolous woman
2.	a-femboy
3.	a person who shares popular misconceptions
4.	a lame or mischievous person
	COVIDIOTS
1.	any of the old world scottish precociously elected officers and pensioners
2.	a person who behaves in an unfriendly and cowardly manner
3.	a person who believes that their tastes or behaviour are superior to those of other people
4.	a person who is secretly willing to obey others
	PLANDEMIC
1.	of or relating to plandelia
2.	an outbreak of a plan demic
3.	a period of plandisone
4.	an act of spreading plandisone

Table 9: Emerging target words and incorrectly generated definitions

In sum, when `T5-base-DG` is presented with new usages from $t^+ > t$, it may (1) produce definitions that are semantically related to an older sense of the target word, (2) rely largely on the context words to generate a definition, (3) adapt the polarity reflected in the usage, (4) generate new made-up words, or (5) use self-reference.

7. Discussion and Conclusion

A language model is capable of temporal generalization if its performance does not deteriorate in a temporally misaligned setup compared to the standard temporally aligned scenario. Lexico-semantic

[14] https://www.dictionary.com/e/slang/simp/

[15] https://www.dictionary.com/e/gender-sexuality/femboy/; https://www.urbandictionary.com/define.php?term=femboy&page=9

temporal generalization in particular refers to a language model's capacity to generalize its semantic knowledge well to data from beyond the training period.

In this work, we fine-tuned T5-base, a Transformer-based language model, for the task of contextualized word definition generation, and tested it on a diagnostic dataset of 400 $\langle word, context\rangle$ pairs from two time periods. These test items concern either semantically stable, semantically changing, or emerging target words. Our hypotheses were that (1) performance on the contextualized word definition generation task would deteriorate for all three target word categories, (2) this deterioration would be more prevalent for words that underwent semantic change in a time period that follows the model's training period, and (3) performance on emerging target words would be lower than for stable target words.

We tested the model on each of the target word categories—stable, changing, and emerging—with input from a temporally aligned and a temporally misaligned time period. If the model were capable of temporal generalization, its performance should not deteriorate between these two time periods. However, our results show an overall performance deterioration of 19.2% for T5-base on the task of contextualized definition generation. Definition quality deteriorates drastically more for changing target words (36.7%) compared to the stable target words (7.5%) and definitions of emerging words were correct only 8.7% of the time, compared to an accuracy of 63.75% for the stable target words. Taken together, these results indicate T5-base lacks strong temporal generalization abilities. Its lexico-semantic understanding is negatively affected by diachronic lexical semantic change and it suffers particularly when shifts in word usage distributions are more marked, as it is the case for emerging words. Our three hypotheses are thus confirmed.

To understand whether a model's inadequate lexico-semantic temporal generalisation would also be revealed by standard language model evaluation metrics, we compared the performance of T5-base on contextualized definition generation to its cross-entropy loss and perplexity scores calculated on the same input examples. Our analysis revealed that cross-entropy loss and perplexity are not consistently reliable indicators of lexico-semantic temporal generalization. Notably, instances with high perplexity scores may still yield accurate definitions, while examples with low perplexity do not necessarily result in accurate definitions. This puts into question the ability of standard evaluation metrics to detect non-trivial temporal generalization failures, which has been posited in related studies (Lazaridou et al. 2021, Ishihara et al. 2022). When reporting performance deterioration, we encourage LM developers and modelers to integrate metrics such as perplexity with more fine-grained evaluations.

7.1 Failure Modes

Our qualitative analysis of the generated definitions sheds light on different types of failures. In some cases, when presented with novel word usages, T5-base outputs definitions that are semantically similar to the original word sense. In other cases, the model shows high sensitivity to usage examples, resulting in over-specific and untruthful definitions. When presented with emerging words, T5-base is more likely to output sentences containing neologisms, content that relates to the polarity that the word usage context conveys, and self-referential language use. Interestingly, most of these generated definitions were fluent, while the factual information that they convey is incorrect. A non-careful reader may in some cases be deceived by such hallucinations (Mickus et al. 2019). This is particularly problematic for practical applications of definition generation models: for example, users who use such a tool to look up the meaning of words which they do not know, would be unable to verify whether the output is correct.

Beyond the temporal generalisation failures on temporally shifted data, we also observe that model performance on changing target words from the time period corresponding to the training data is substantially lower than for the stable words in the same time period. At first sight, this is an unexpected result, as usage examples for both word categories originate from a time period on which the model was trained. One possible explanation is that these changing words were *already* unstable during that time period. This aligns with empirical evidence that words undergoing semantic change typically go through polysemous stages before a dominant sense is established—the so-called '*the Law of Innovation*' (Tahmasebi et al. 2021). Alternatively, these findings can be explained by the fact that the changing target words are relatively infrequent in the training time period compared to an average stable word in the English vocabulary. This would negatively impact understanding of unstable words in two ways. First, as a practical consequence, the pre-trained model `T5-base` is likely exposed to fewer training instances for these changing words in the first place, resulting in their representations to be of lower quality. Second, frequent words are known to change more slowly than infrequent words—a phenomenon referred to as the *Law of Conformity* (Tahmasebi et al. 2021), and the quality of their representations may benefit from their stable usage.

7.2 Limitations

To fully understand the scope of our findings, especially in relation to current trends in NLP, it is important to discuss a few limitations of our experimental setup.

Transferability of findings to other language models Our experiments were conducted on a single language model, so it is natural to question whether they would transfer to models with different architectures and pre-training data. It is possible that other pre-trained LLMs would not show the same degree of performance deterioration on the contextualized definition generation task. More recent modern models such as GPT-3 (Brown et al. 2020), BART (Lewis et al. 2020), LLAMA (Touvron et al. 2023) have been shown to outperform `T5` in several scenarios and they might be expected to be more robust to performance deterioration due to temporal misalignment. Still, these more modern LLMs make use of similar Transformer architectures and are optimized via standard language modelling training objectives. There are, therefore, no principled reasons to assume that they more robustly generalise to temporally shifted data beyond the fact that their larger training datasets might be more likely to cover word usages similar to those in the test set—which would mean that weaker performance deterioration could be hardly considered a case of generalization (Hupkes et al. 2023).

Fine-tuning as a testing interface We proposed assessing the lexico-semantic temporal generalization of a model using the definition generation task and we chose fine-tuning as a way to instruct the model to perform it. This implies the temporal generalization test is conducted on a model checkpoint that slightly differs from the target model. While the impact of our definition generation fine-tuning on the model parameters is likely very low relative to the full pre-training phase, we did not test whether adaptation resulted in the loss of prior knowledge. Alternative approaches that are less prone to this issue could be based on in-context task learning, with the language model being shown a few examples of the definition generation task at inference time, without any parameter update. Recent results on zero-shot definition generation (Giulianelli et al. 2023) and few-shot word-in-context tasks (Periti et al. 2024) suggest these approaches might be suitable only for the largest and most recent language models.

Challenge sets and their scalability Our experimental setup is limited to 400 $\langle word, context\rangle$ pairs in total, 20 target words per category, and two time periods, due to the onus of collecting human annotations for the model generations. The reliance on human annotations can in this sense be considered a limitation of our approach, as it limits its scalability. Automatic evaluation of the generated definitions would in principle be possible if the challenge set consisted of diachronic dictionary data, i.e., a set of timestamped definitions of target words together with time-specific usage examples—the English Urban Dictionary could for instance be a possible source of data for this. However, automatic NLG metrics such as BLEU, BERTScore, or NIST have not been validated against human definition quality judgements and are likely to produce evaluation scores biased by stylistic mismatch between dictionary entries and generated definitions. In any case, our proposed task is not merely meant as a substitute for other quantitative measures of temporal generalization. Our experiments provide us with targeted human interpretable assessments, which, even if conducted at a small scale, can provide valuable qualitative insights. An aspect of our setup where human annotation may have been appropriate is the selection of the example usages for the target words. In this case, to avoid time-consuming manual inspection, we opted for selecting the example sentences randomly, in contrast to the example usages from the Oxford data set on which the model was fine-tuned, which were handpicked by expert lexicographers. This random selection is not optimal, because the context sentences may contain ambiguous usages of the target words, and there is no guarantee that the sentences sampled from C_2 actually display new usages. This is a limitation of our automatic approach, and we acknowledge that manual selection or filtering of the context sentences may have led to more robust semantic change detection, albeit to a methodology that is less scalable.

7.3 Outlook

Our findings regarding the fact that lack of temporal generalization relates to semantic change, risking bias and hallucination, could carry over to other tasks such as summarization, question answering, or translation, among others. If LLMs struggle more to generate accurate definitions for input where semantic change is present, it is likely that other generative tasks are also negatively affected over time due to semantic change, which could have a real impact on society through practical applications. Notably, the proposed methodology of semantic assessment using the task of contextualized word definition generation could also be used to examine lexico-semantic generalization in types of domain shift that are not temporal. For instance, future work could use our semantic assessment task to examine whether a model trained on news data is able to generalize well to academic papers. Overall, the contextualized definition generation task is a promising semantic assessment, demonstrating that performance deterioration is stronger for semantically changing words, and providing interpretable insight into the possible mistakes that a model may perpetrate due to meaning shift.

Acknowledgements

This paper builds upon the preliminary work presented by Luden (2023). We are grateful to Jelke Bloem and Pia Sommerauer for their very useful comments on the preliminary version. We would also like to thank Andrey Kutuzov and Pauline Sander for their feedback throughout the project. This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

References

Agarwal, Oshin and Ani Nenkova (2022), Temporal effects on pre-trained models for language processing tasks, *Transactions of the Association for Computational Linguistics* **10**, pp. 904–921, MIT Press, Cambridge, MA. https://aclanthology.org/2022.tacl-1.53.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn (2020), The Pushshift Reddit dataset, *Proceedings of the International AAAI Conference on Web and Social Media* **14**, pp. 830–839.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021), On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. https://doi.org/10.1145/3442188.3445922.

Biesialska, Magdalena, Katarzyna Biesialska, and Marta R. Costa-jussà (2020), Continual lifelong learning in natural language processing: A survey, *in* Scott, Donia, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 6523–6541. https://aclanthology.org/2020.coling-main.574.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020), Language models are few-shot learners, *in* Larochelle, H., M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, Curran Associates, Inc., pp. 1877–1901.

Chen, Shuguang, Leonardo Neves, and Thamar Solorio (2021), Mitigating temporal-drift: A simple approach to keep ner models crisp, *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pp. 163–169. https://aclanthology.org/2021.socialnlp-1.14.

Dhingra, Bhuwan, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen (2022), Time-aware language models as temporal knowledge bases, *Transactions of the Association for Computational Linguistics* **10**, pp. 257–273, MIT Press, Cambridge, MA. https://aclanthology.org/2022.tacl-1.15.

Eisenschlos, Julian Martin, Jeremy R. Cole, Fangyu Liu, and William W. Cohen (2023), WinoDict: Probing language models for in-context word acquisition, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Dubrovnik, Croatia, pp. 94–102. https://aclanthology.org/2023.eacl-main.7.

Fellbaum, Christiane (1998), *WordNet: An Electronic Lexical Database*, Bradford Books. https://mitpress.mit.edu/9780262561167/.

Gadetsky, Artyom, Ilya Yakubovskiy, and Dmitry Vetrov (2018), Conditional generators of words definitions, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Melbourne, Australia, pp. 266–271. https://aclanthology.org/P18-2043.

Giulianelli, Mario, Iris Luden, Raquel Fernández, and Andrey Kutuzov (2023), Interpretable word sense representations via definition generation: The case of semantic change analysis, *in* Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, pp. 3130–3148. https://aclanthology.org/2023.acl-long.176.

Huang, Han, Tomoyuki Kajiwara, and Yuki Arase (2021), Definition modelling for appropriate specificity, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 2499–2509. https://aclanthology.org/2021.emnlp-main.194.

Hupkes, Dieuwke, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin (2023), A taxonomy and review of generalization research in nlp, *Nature Machine Intelligence* **5** (1010), pp. 1161–1174, Nature Publishing Group.

Ishihara, Shotaro, Hiromu Takahashi, and Hono Shirai (2022), Semantic Shift Stability: Efficient Way to Detect Performance Degradation of Word Embeddings and Pre-trained Language Models, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics, Online only, p. 205–216. https://aclanthology.org/2022.aacl-main.17.

Jang, Joel, 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* Goldberg, Yoav, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 6237–6250. https://aclanthology.org/2022.emnlp-main.418.

Lazaridou, Angeliki, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. (2021), Mind the gap: Assessing temporal generalization in neural language models, *Advances in Neural Information Processing Systems* **34**, pp. 29348–29363. https://proceedings.neurips.cc/paper/2021/hash/f5bf0ba0a17ef18f9607774722f5698c-Abstract.html.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (2020), BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 7871–7880. https://aclanthology.org/2020.acl-main.703.

Loper, Edward and Steven Bird (2002), NLTK: The Natural Language Toolkit, *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, Association for Computational Linguistics, USA, p. 63–70. https://doi.org/10.3115/1118108.1118117.

Loureiro, Daniel, Aminette D'Souza, Areej Nasser Muhajab, Isabella A. White, Gabriel Wong, Luis Espinosa-Anke, Leonardo Neves, Francesco Barbieri, and Jose Camacho-Collados (2022a), TempoWiC: An evaluation benchmark for detecting meaning shift in social media, *Proceedings of the 29th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pp. 3353–3359. https://aclanthology.org/2022.coling-1.296.

Loureiro, Daniel, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados (2022b), TimeLMs: Diachronic language models from Twitter, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Dublin, Ireland, pp. 251–260. https://aclanthology.org/2022.acl-demo.25.

Luden, Iris (2023), Beyond perplexity: Examining temporal generalization of large language models via definition generation, *Unpublished Master's thesis, University of Amsterdam, Amsterdam.* https://eprints.illc.uva.nl/id/eprint/2291/1/MoL-2023-33.text.pdf.

Luu, Kelvin, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith (2022), Time Waits for No One! Analysis and Challenges of Temporal Misalignment, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Seattle, United States, pp. 5944–5958. https://aclanthology.org/2022.naacl-main.435.

Mickus, Timothee, Denis Paperno, and Mathieu Constant (2019), Mark my Word: A Sequence-to-Sequence Approach to Definition Modeling, *DL4NLP 2019* p. 1.

Mickus, Timothee, Denis Paperno, and Mathieu Constant (2022), How to dissect a Muppet: The structure of transformer embedding spaces, *Transactions of the Association for Computational Linguistics* **10**, pp. 981–996, MIT Press, Cambridge, MA. https://aclanthology.org/2022.tacl-1.57.

Noraset, Thanapon, Chen Liang, Larry Birnbaum, and Doug Downey (2017), Definition modeling: Learning to define word embeddings in natural language, *Thirty-First AAAI Conference on Artificial Intelligence.*

Periti, Francesco, Haim Dubossarsky, and Nina Tahmasebi (2024), (Chat) GPT v BERT: Dawn of justice for semantic change detection, *arXiv preprint arXiv:2401.14040.*

Raffel, Colin, 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, *J. Mach. Learn. Res.*, JMLR.org.

Ranjan, Nihar, Kaushal Mundada, Kunal Phaltane, and Saim Ahmad (2016), A survey on techniques in NLP, *International Journal of Computer Applications* **134** (8), pp. 6–9. https://doi.org/10.5120/ijca2016907355.

Rosin, Guy D., Ido Guy, and Kira Radinsky (2022), Time masking for temporal language models, *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, Association for Computing Machinery, New York, NY, USA, p. 833–841.

Röttger, Paul and Janet Pierrehumbert (2021), Temporal adaptation of BERT and performance on downstream document classification: Insights from social media, *Findings of the Association for Computational Linguistics: EMNLP 2021*, Association for Computational Linguistics, Punta Cana, Dominican Republic, pp. 2400–2412. https://aclanthology.org/2021.findings-emnlp.206.

Salazar, Julian, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff (2020), Masked language model scoring, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.acl-main.240.

Schlechtweg, Dominik, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde (2019), A wind of change: Detecting and evaluating lexical semantic change across times and domains, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 732–746. https://aclanthology.org/P19-1072.

Sinha, Koustuv, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela (2021), Masked language modeling and the distributional hypothesis: Order word matters pre-training for little, *in* Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 2888–2913. https://aclanthology.org/2021.emnlp-main.230.

Su, Zhaochen, Zecheng Tang, Xinyan Guan, Lijun Wu, Min Zhang, and Juntao Li (2022), Improving temporal generalization of pre-trained language models with lexical semantic change, pp. 6380–6393, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates. https://aclanthology.org/2022.emnlp-main.428.

Tahmasebi, Nina, Lars Borin, and Adam Jatowt (2021), *Survey of computational approaches to lexical semantic change detection*, Language Science Press, Berlin, p. 1–91.

Touvron, Hugo, 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 (2023), Llama: Open and efficient foundation language models. cite arxiv:2302.13971. http://arxiv.org/abs/2302.13971.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin (2017), Attention is all you need, *Advances in Neural Information Processing Systems*, Curran Associates, Inc.

Appendix A. GenBench evaluation card

<table>
<tr><th colspan="12">Motivation</th></tr>
<tr><td colspan="3"><i>Practical</i></td><td colspan="3"><i>Cognitive</i></td><td colspan="3"><i>Intrinsic</i>
□</td><td colspan="3"><i>Fairness</i></td></tr>
<tr><th colspan="12">Generalisation type</th></tr>
<tr><td colspan="2"><i>Compositional</i></td><td colspan="2"><i>Structural</i></td><td colspan="2"><i>Cross Task</i></td><td colspan="2"><i>Cross Language</i></td><td colspan="2"><i>Cross Domain</i>
□</td><td colspan="2"><i>Robustness</i></td></tr>
<tr><th colspan="12">Shift type</th></tr>
<tr><td colspan="3"><i>Covariate</i></td><td colspan="3"><i>Label</i></td><td colspan="3"><i>Full</i>
□</td><td colspan="3"><i>Assumed</i></td></tr>
<tr><th colspan="12">Shift source</th></tr>
<tr><td colspan="3"><i>Naturally occuring</i>
□</td><td colspan="3"><i>Partitioned natural</i></td><td colspan="3"><i>Generated shift</i></td><td colspan="3"><i>Fully generated</i></td></tr>
<tr><th colspan="12">Shift locus</th></tr>
<tr><td colspan="3"><i>Train–test</i></td><td colspan="3"><i>Finetune train–test</i>
□</td><td colspan="3"><i>Pretrain–train</i></td><td colspan="3"><i>Pretrain–test</i></td></tr>
</table>

Appendix B. Annotation guidelines

You are provided with a spreadsheet with four columns: **Targets**, **Judgement**, **Example** and **Definition**. In every row, there is one English target word in the **Target** column, one example sentence in which this target word is used in the **Example** column, and one definition sentence or phrase in the **Definition** column. The definition has been generated by a large language model and it is a context-specific definition for the target word in the example sentence.

Words can have different meanings, depending on the context in which they are used. The possible meanings that a word in different contexts can have are called *senses.* A popular example is the word **bank**, which is a polysemous word:

Sentence 1: I need to get some money from the *bank*
Sentence 2: I'm walking along the river *bank.*

In sentence 1, the sense of the target word *bank* can be defined as "An institution that invests money deposited by customers or subscribers". In sentence 2, on the other hand, the target word *bank* refers to the sense that can be defined as "the sloping, vertical, or overhanging edge of a river or other watercourse".

Your task is to judge for each row whether the definition of the target word in the example sentence is correct. That is, the definitions must be:

- **Truthful**: i.e. should reflect exactly the sense in which the target word is occurring in the example sentence. Ideally, the definition should be specific enough so as not to mix with other senses, while general enough so as not to describe information of the example sentence that does not concern the target word.
- **Fluent** i.e., feeling like natural English sentence or phrase, without grammar errors, utterances broken mid-word, etc.

Task instructions

You have to fill in the **Judgements** column with one of five values:

0: The definition is incorrect; not truthful or not fluent

1: The definition is partially incorrect; it is either not truthful or not fluent, but it does reflect some information related to the sense of the target word in the example sentence

2: The definition is mostly correct; it is truthful and fluent, but could be better nuanced

3: The definition is correct.

-10: The definition is self-referential; i.e. refers back to the target word itself.

Example

The word *dress* can be used as a noun, or as a verb. Consider the following pairs of example sentences and (correct)definitions:

A Target word: *dress*
Example sentence: I am wearing my beautiful pink *dress*.
Definition: a one-piece garment, typically extending down over the legs in a skirt [16]

B Target word: *dress*
Example sentence: I want to *dress* up nicely for the party.
Definition: to clothe oneself

The definitions above are correct; they are fluent and truthful, and therefore you would judge them with a **3** in the 'judgements' column. If, however, the definition of B would be provided for the example sentence of A (or vice versa), the definitions would be *incorrect* for the target word in the example sentence, because it defines the wrong sense of the target word. In this case, you would judge with the score **0**.

Too specific or too general

Definitions can be too specific or too general to the context in which it is used. An example of a too specific definition for Example sentence A is:

a one-piece garment that is pink and beautiful

[16] The complete definition from the Online Oxford English Dictionary is:(`https://www.oed.com/search?searchType=dictionary&q=dress&_searchBtn=Search`)

This definition is too specific because the sense of the target word *dress* does not necessarily require the dress to be pink nor beautiful, the adjectives in this sentence only specify what the color of the dress is. You should judge this with a **1**.

An example of a too general definition for sentence 4 would be:

> a piece of cloth

This definition is too general for this example sentence, because 'a piece of cloth' can also describe many other objects, like a t-shirt or a towel, which are not possible in this sentence. You should judge the generated definitions with a **1**, as the definition is not sufficiently truthful.

Definitions could be fluent and truthful, but could be better nuanced, for example:

> a one-piece clothing, often worn by women and girls

This definition is truthful and fluent, and undoubtedly refers to the correct sense of the target word. However, it might be improved with some extra nuance or information. Therefore you would judge this definition by a **2**.

Self-reference

When self-reference occurs, the definition is considered *incorrect* and should receive the special label **-10**. An example of a self-referential definition is:

> Target word: self-conscious
> Definition: the state of being self-conscious

Appendix C. Top 20 trending words

	Trending Word	CD
1	pandemic	0.90
2	quarantine	0.88
3	vaccine	0.45
4	vaccinated	0.53
5	lockdown	0.96
6	moots	0.37
7	corona	0.98
8	distancing	0.83
9	vaccination	0.54
10	virus	0.52
11	airdrop	0.52
12	yacht	0.72
13	masks	0.60
14	ukraine	0.35
15	vaccines	0.42
16	mandates	0.60
17	ukrainian	0.43
18	doge	0.51
19	staking	0.61
20	bodybuilding	0.68

Table 10: Top 20 trending words and their cosine distance scores.

Appendix D. Alternative judgement aggregations

Besides taking the ‘majority vote’ to aggregate judgements, we also computed the consensus vote. In consensus voting, a definition is considered correct whenever *all (three) annotators judged it to be correct.* Recall that judgement scores of 2 and 3 are considered correct, while the judgements -10, 0, 1 are considered incorrect. Aggregating the judgements by consensus voting generally displays the same trends as in majority voting (see Table 4), except that the stable target words of C_2 are judged to better than those of C_1. This would imply that the performance of `T5-base` does not deteriorate for input that concerns stable target words.

Category	C_1	C_2	$C_1 \cup C_2$
stable	**43.75%**	**47.50%**	**45.62%**
changing	**27.50%**	12.50%	20.00%
stable + changing	**35.625%**	30.0%	32.81%
emerging	-	2.50%	-
total	**35.625%**	20.83%	26.75%

Table 11: Accuracy according to consensus vote