Automatically Generated Definitions and their utility for Modeling Word Meaning

Anonymous ACL submission

Abstract

Modeling lexical semantics is a challenging task, often suffering from interpretability pitfalls. In this paper, we delve into the generation of dictionary-like sense definitions and explore their utility for modeling word meaning. We fine-tuned two Llama models and include an existing T5-based model in our evaluation. Firstly, we evaluate the quality of the generated definitions on existing benchmarks, setting new state-of-the-art results for the Definition Generation task. Next, we explore the use of definitions generated by our models as intermediate representations subsequently encoded as sentence embeddings. We evaluate this approach on lexical semantics tasks such as the Word-in-Context, Word Sense Induction, and Lexical Semantic Change, setting new state-of-the-art results in all three tasks.

1 Introduction

Modeling *lexical semantics* using unstructured text has a longstanding history in Natural Language Processing due to its crucial role in both Natural Language Understanding and Natural Language Generation (Karanikolas et al., 2024; Pustejovsky and Boguraev, 1993). Over the past decades, there have been many relevant technological developments: from count-based (Naseem et al., 2021) to static (Mikolov et al., 2013) and contextualized (Peters et al., 2018) language models, and most recently, generative models (Hadi et al., 2023). Each of these advancements has contributed significantly to the goal of *modeling the meaning of words*.

Modern language models are based on the Transformer (Vaswani et al., 2017) architecture. Given a word, these models generate semantic representations for each occurrence of the word based on its surrounding context (Apidianaki, 2023). Ideally, these representations should be similar for semantically related word usages and different for semantically distinct ones. Typically, *contextualized* vectors (i.e., embeddings, Pilehvar and Camacho-Collados, 2021) or lexical substitutes (i.e., bag-of-words, Arefyev and Zhikov, 2020) are employed to represent word usages. However, recent advancements in text generation are shifting the attention towards representing word usages through generated *sense definitions* (Giulianelli et al., 2023).

Automatically generated sense definitions provide a dual advantage. Firstly, they distill the information stored in a sentence by abstracting away from the context. Their use potentially condenses various word usage representations pertaining to the same underlying meaning. Secondly, generated definitions provide a means to directly interpret word meaning from unstructured text, thereby enabling language models to serve as surrogate for dictionaries when encountering unfamiliar words (Malkin et al., 2021), or known words in unfamiliar settings (Weiland et al., 2023).

In this work, we automatically generate definitions for words *in-context* by relying on two fine-tuned variants of the Llama chat models (Touvron et al., 2023) refined through instruction tuning (Zhang et al., 2024) on lexicographic resources. We call the models `LlamaDictionary` and assess their performance in Definition Generation, achieving new state-of-the-art results on multiple datasets.

We further extend our evaluation by using `LlamaDictionary` and the existing `Flan-T5` model fine-tuned by Giulianelli et al. (2023) for large scale modeling of word meaning. Specifically, we employ the generated sense definitions as intermediate sense representations. These representations are encoded using a pretrained sequence embedding model rather than using standard token embeddings. We evaluate our approach on three popular Natural Language Processing tasks, namely Word-in-Context, Word Sense Induction, and Lexical Semantic Change, achieving new state-of-the-art results on all three tasks.

Our original contribution:

- We introduce `LlamaDictionary`, a novel fine-tuned large language model designed to generate sense definitions for words *in-context*.
- We evaluate the use of `LlamaDictionary` and existing `Flan-T5` with thirteen `SBERT` models, achieving new state-of-the-art results in the Definition Generation task.
- We demonstrate the effectiveness of `LlamaDictionary` and `Flan-T5` as a preprocessing tool for large-scale word meaning analysis and achieve state-of-the-art results in the Word-in-Context, Word Sense Induction, and Lexical Semantic Change task.

2 Background and related work

2.1 Word usage representations

With the advent of Transformers, we have witnessed the emergence of large language models capable of contextualizing words within diverse contexts. Unlike static models (Pennington et al., 2014), we now rely on a multitude of contextualized embeddings per word. On one hand, this capability represents an invaluable tool for modeling lexical semantics (Petersen and Potts, 2023), as distances between embeddings have proven to be excellent discriminators of word meaning. On the other hand, it poses interpretability challenges, as embeddings tend to represent contextual variance rather than lexicographic senses (Kutuzov et al., 2022). Further challenges arise from the broad and heterogeneous distribution of semantic structure across embedding dimensions (Senel et al., 2018).

Lexical substitutes are often employed as alternative representations to raw embeddings (Alagic et al., 2018). These representations consist of sets of automatically generated replacements for specific occurrences of words in-context. Unlike embeddings, lexical substitutes can be directly inspected to infer word meaning. However, the interpretation process requires more time and effort compared to the conventional practice of consulting a dictionary for satisfying meaning definitions. Additionally, interpreting the meaning of a word remains challenging, as lexical substitutes can include stopwords and partial word pieces (Card, 2023), equally plausible alternatives with different meanings (Chiang and Lee, 2023), and even contradictory replacements (Justeson and Katz, 1991).

With the recent advancements in text generation, *automatically generated sense definitions* become a viable approach for word usage representation, as these definitions offer descriptive interpretations of words *in-context*, providing a valuable tool with a level of interpretability comparable to manually curated vocabularies (Gardner et al., 2022).

2.2 Generating word sense definitions

Generating word sense definitions has initially gained attention to enhance the interpretability of static embeddings (Mickus et al., 2022; Gadetsky et al., 2018). Originally, the task involved generating a natural language definition given a single embedding of a target word (Noraset et al., 2017). However, since words can carry multiple meanings, advancements in contextualized modeling have shifted the focus to the generation of appropriate sense definitions for words in context (Zhang et al., 2022; Huang et al., 2021; Mickus et al., 2019; Ishiwatari et al., 2019).

Generated definitions are useful in a multitude of applications such as the generation of lexicographic resources for low-resource languages (Bear and Cook, 2021), explaining register- or domain-specific vocabulary (Ni and Wang, 2017; August et al., 2022), or language learning scenarios (Zhang et al., 2023; Kong et al., 2022; Yuan et al., 2022).

While early works use sequence-to-sequence models for definition modeling (Ni and Wang, 2017; Gadetsky et al., 2018; Mickus et al., 2019), later works utilize pretrained language models such as BART (Bevilacqua et al., 2020; Segonne and Mickus, 2023; Lewis et al., 2020) and T5 (Huang et al., 2021; Tseng et al., 2023; Raffel et al., 2020).

More recently, Giulianelli et al. (2023) has proposed using generated definitions as interpretable word usage representation for the analysis of lexical semantic change and provided a new model called `Flan-T5`. Inspired by this work, we follow the idea that definitions can be used as interpretable representations and also position our work with a focus on modeling word meaning and meaning change. Inspired by Bevilacqua et al. (2020), we encode definitions as sentence embeddings. However, we model the meaning of words *in-context* with a single sense definition rather than a set.

3 Automatic definition generation

In this work, we fine-tuned two popular open-source generative models through instruction tun-

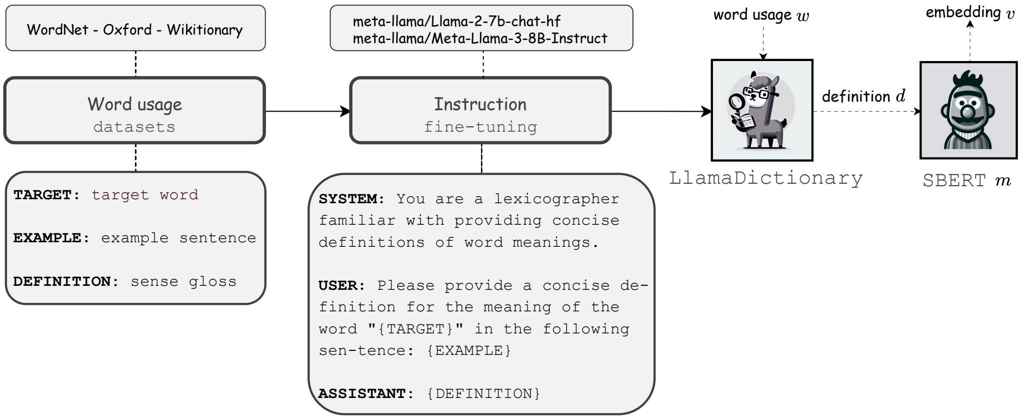
WordNet - Oxford - Wikitionary
meta-llama/Llama-2-7b-chat-hf
meta-llama/Meta-Llama-3-8B-Instruct
word usage w
embedding v
Word usage
datasets
Instruction
fine-tuning
definition d
LlamaDictionary
SBERT m
TARGET: target word
EXAMPLE: example sentence
DEFINITION: sense gloss
SYSTEM: You are a lexicographer familiar with providing concise definitions of word meanings.
USER: Please provide a concise de-finition for the meaning of the word "{TARGET}" in the following sen-tence: {EXAMPLE}
ASSISTANT: {DEFINITION}

Figure 1: `LlamaDictionary` is a Llama chat model fine-tuned with lexicographic resources to generate a sense definition from an input word usage.

ing, namely Llama2chat[1] and Llama3instruct[2]. We specifically chose to fine-tune chat models because they were already optimized to generate responses adhering to specific instruction prompts. We call the models resulting from fine-tuning `LlamaDictionary`. In the following, we refer to `Llama2Dictionary` and `Llama3Dictionary` for the fine-tuned versions of Llama2chat and Llama3instruct, respectively.

Using `Llama2Dictionary` and `Llama3-Dictionary`, we complement the existing Flan-T5 3B model by Giulianelli et al. (2023) with two larger Llama 7B and 8B, chat-based versions.

3.1 Data

We fine-tune Llama2chat and Llama3instruct on the same English data used by Giulianelli et al. (2023). The data consists of *word usages* $\langle w, e, d \rangle$, where w represents a target word, e denotes an example context where w occurs, and d is a human-curated definition for the lexicographic sense of the word w in the example e. The considered word usages span three benchmarks previously extracted from the **Oxford** English Dictionary (Gadetsky et al., 2018), **WordNet** (Ishiwatari et al., 2019), and **Wikitionary** (Mickus et al., 2022), respectively. However, while Giulianelli et al. (2023) use all the Train-Dev-Test partitions during fine-tuning, we use only Train and Dev and reserve Test for evaluation purposes. Table 1 reports the main statistics of these benchmarks.

		Oxford	WordNet	Wikitionary	Tot.
Train	# words	33,128	7,935	18,030	45,070
	# definitions	97,802	13,854	31,142	142,798
	# def. per word	2.95	1.75	1.73	3.17
Dev	# words	8,863	998	2,561	11,666
	# definitions	12,222	1,748	4,525	18,495
	# def. per word	1.38	1.75	1.77	1.59
Test	# words	8,848	1,001	2,361	11,718
	# definitions	12,228	1,774	4,436	18,438
	# def. per word	1.38	1.77	1.69	1.57

Table 1: Train-Dev-Test partitions of the considered benchmarks. For each partition, we report the number of unique words, the number of unique definitions, and the average number of definitions per target word.

3.2 Fine-tuning

Llama2chat and Llama3instruct with 7 and 8 billion parameters, respectively, are large, decoder-only architectures trained on publicly available online data, followed by supervised fine-tuning through instruction tuning (Zhang et al., 2024) and iterative refinement using reinforcement learning from human feedback (Kaufmann et al., 2024). We further fine-tuned these models through instruction tuning for sense definition generations.

Given the high costs associated with fine-tuning large language models, we employed a parameter-efficient fine-tuning (Han et al., 2024) that enables efficient adaptation by only fine-tuning a small number of additional model parameters instead of the entire model. This approach significantly reduces computational and storage costs. Specifically, we fine-tuned using Low-rank Adapter (LoRA, Hu et al., 2021). [3] Experimented hyper-

[1] `meta-llama/Llama-2-7b-chat-hf`

[2] `meta-llama/Meta-Llama-3-8B-Instruct`

[3] We have also experimented with Quantization combined with LoRA (QLORA, Dettmers et al., 2023) obtaining very similar evaluation results (see Figure 4). These are omitted due to space restriction but will be available in our **Github** repository where we will publish all our code, data, and results.

parameters are reported in Table 10 and 11.

For fine-tuning, we used cross-entropy loss calculated on all tokens over 4 epochs, with a batch size of 32, a maximum sequence length of 512, and *packing* to train efficiently on multiple samples simultaneously (Kosec et al., 2021).

In line with Huerta-Enochian (2024), who demonstrated that prompt loss can be safely ignored for many datasets, we observed lower preliminary results in the evaluation tasks for models chosen based on validation performance. Therefore, we selected the final model based on the checkpoint at the last training epoch.

3.3 Instruction-tuning

We fine-tuned Llama2chat and Llama3instruct using the prompt shown in Figure 1. For each word usage $\langle w, e, d\rangle$, we substituted `TARGET` with the actual target w, and `EXAMPLE` and `DEFINITION` with the example e and the definition d, respectively.

For our prompt, we drew inspiration from prompts used in previous work, specifically, we employed a prompt similar to those used by Giulianelli et al. (2023). In line with Li et al. (2023), we incorporated an emotional stimulus (in Figure 1, `Please`) to enhance the performance. Additionally, similarly to Kocoń et al. (2023); Laskar et al. (2023); Periti et al. (2024b), we structured our prompt in a format that facilitates parsing and comprehension.

4 Evaluation setup

Our evaluation is structured into two parts. First, we assess the quality of definitions generated by `LlamaDictionary` and `Flan-T5` through the Definition Generation (DG) task. For this evaluation, we directly utilize the generated sense definitions.

Next, we explore their utility in three popular Natural Language Processing tasks, namely Word-in-Context (WiC), Lexical Semantic Change (LSC), and Word Sense Induction (WSI). Specifically, instead of using standard token embeddings, we view sense definitions as intermediate sense representations and encode these as embeddings through a pretrained sequence embedding model. Formally, this means that: given an occurrence of a word w, we employ a generative model g (i.e., `LlamaDictionary` or `Flan-T5`) to generate a definition d, which we subsequently encode as a vector v using a sentence embedding model m, i.e., $v = m(d) = m(g(w))$.

Following Giulianelli et al. (2023), we used the `all-distilroberta-v1` sentence SBERT model (Reimers and Gurevych, 2019) to encode definitions as contextualized sentence embeddings. To validate our results, we also evaluate twelve other SBERT models which show comparable results. Furthermore, we extend our evaluation by also considering generated definitions by the `Flan-T5` model recently fine-tuned by Giulianelli et al. (2023)[4] as this model has not been evaluated on the WiC, WSI, and LSC tasks previously.

4.1 Definition generation (DG)

> Given a target word w and an example usage e, the task is to generate a natural language definition d that is grammatical, fluent, and faithful to the meaning of the target word w as used in the example usage e (Giulianelli et al., 2020).

We assess the models in generating sense definitions for both familiar (*Seen* during training) and unfamiliar (*Unseen*) domains and styles.

For *Seen* evaluation, we use the **WordNet**, **Oxford**, and **Wiktionary** Test sets (see Table 1).

For *Unseen* evaluation, we consider the Test sets of two additional benchmarks comprising word usages from The **Urban** Dictionary (the largest online slang dictionary) (Ni and Wang, 2017) and **Wikipedia** (with rare words and phrases) (Ishiwatari et al., 2019). The Train set of these benchmarks were not considered during training.

The decision to exclude **Urban** and **Wikipedia** from training was threefold. Firstly, their exclusion broadens the scope of our evaluation by considering familiar and unfamiliar usages. Secondly, it enabled a direct comparison with `Flan-T5`, a T5-based (Raffel et al., 2020) model. Finally, we refrained from fine-tuning the model with bad, slang, or offensive words, and with numerous erroneous entries (e.g., definitions comprising single Arabic numerals or part-of-speech tags) in **Urban** (Huang et al., 2021). Table 3 reports the main statistics of these benchmarks.

For comparison with previous work, we evaluated `LlamaDictionary` and `Flan-T5` by considering standard Natural Language Generation metrics such as BLEU (Papineni et al., 2002), NIST (Doddington, 2002), SacreBLEU (Post, 2018), ROUGE-

Code and data are submitted as supplementary material.

[4] `ltg/flan-t5-definition-en-xl`

Target w	Example e	Definition d	`LlamaDictionary`
revitalize	This food revitalized the patient	Restore strength	Give new life or energy to
glove	Maxwell gloved his hand so that he would n't leave fingerprints , then pulled the trigger	To put a glove or gloves on .	Wear a glove to protect the hand when performing an activity

Table 2: Examples of pertinent definitions generated by `LlamaDictionary` for two word usages. The generated definitions are unfairly penalized by standard evaluation metrics.

		Urban	Wikipedia
Test	# words	25,909	56,008
	# definitions	34,974	8,193
	# def. per word	1.35	6.84

Table 3: Test partitions of *Unseen* DG benchmarks.

L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and EXACT MATCH. Since some pertinent definitions may be unfairly penalized due to missing lexical overlap (see Table 2), we follow Giulianelli et al. (2023) and consider BERT-F1 Score (Zhang et al., 2020), which represents a semantic and thus valuable metric for this task.

4.2 Word-in-Context (WiC)

> Given a target word w and two contexts c_1 and c_2 where w occurs, the task is to identify whether the occurrences of w in c_1 and c_2 correspond to the same meaning or not (Pilehvar and Camacho-Collados, 2019).

We evaluate the utility of sense definitions using sequence embeddings $v = m(g(w))$ on the original WiC benchmark (Pilehvar and Camacho-Collados, 2019). We refrain from using the Train set and instead generate two embeddings, v, for each context pair (one for c_1 and one for c_2) within the Dev and Test partitions (see Table 4). To address the WiC task, we then train a threshold-based classifier, for each tested model, using the cosine distance between the two embeddings of each pair in the Dev set. The training process involves selecting the threshold that maximizes the performance on the Dev set. Finally, we apply this classifier to conduct our evaluation over the Test set. We utilize accuracy as the assessment metric for comparison with previous work (Pilehvar and Camacho-Collados, 2019).

	WiC	
Partition	**Dev**	**Test**
# pairs	638	1,400
# words	599	1,184

Table 4: Test-Dev partitions for Word-in-Context.

4.3 Lexical Semantic Change (LSC)

> Given a set of target words w and two corpora C_1 and C_2 of different time periods, the task is to rank the targets according to their degree of *lexical semantic change*[a] between C_1 and C_2 (Schlechtweg et al., 2020).
>
> [a] "Innovations which change the lexical meaning rather than the grammatical function of a form" (Bloomfield, 1933)

We evaluate our approach on the original SemEval-English LSC benchmark (Schlechtweg et al., 2020). The dataset consists of two corpora and a test set of 46 target words (see Table 5). Train and Dev sets are not available as the task is set in an unsupervised scenario. To address the LSC task, we leverage popular methods generally applied using word embeddings rather than sentence embeddings (Periti and Tahmasebi, 2024). In particular, we evaluate two different approaches:

Average Pairwise Distance (APD) is defined as *form-based* method, meaning that it quantifies change without modeling the underlying meanings of the words. Given a word w, APD computes the degree of change as the average pairwise distance between the embeddings of w generated for C_1 and C_2 (Giulianelli et al., 2020).

Average Pairwise Distance Between Sense Prototypes (APDP) is defined as *sense-based* method, meaning that it quantifies change after modeling the underlying meanings of the words via clustering. Following previous work (Rother et al., 2020) and the recent BERTopic pipeline (Grootendorst, 2022), we consider the HDBSCAN algorithm (McInnes et al., 2017). Given a word w, APDP computes the degree of change as the average pairwise distances between the sense prototypes of w in the time periods C_1 and C_2, where sense prototypes are the set of embeddings obtained by averaging the embeddings of C_1 and C_2 in each cluster, respectively (Kashleva et al., 2022).

For comparison with previous work, we utilize the Spearman rank correlation between gold scores and predictions as the assessment metric.

Test	LSC - WSI
# words	46
# clusters per word	9.4
max # of clusters	55
min # of clusters	1

Table 5: Test set for Lexical Semantic Change and Word Sense Induction, EN portion of SemEval-2020 Task 1.

4.4 Word Sense Induction (WSI)

> Given a set of occurrences for a target word w, the task is to automatically determine the different senses of w without relying on predefined sense inventories (Agirre and Soroa, 2007).

For simplicity, we follow the recent comparison by Periti and Tahmasebi (2024) and perform a WSI evaluation on the same benchmark used for the LSC evaluation, as it also includes gold scores for WSI. Thus, we evaluate the clustering result obtained by using HDBSCAN against labels provided for clusters in the LSC data.

As assessment metrics, we utilize Rand Index (RI) (Rand, 1971) and its Adjusted version (ARI) (Hubert and Arabie, 1985) as well as Purity (Manning, 2009). RI/ARI evaluate the similarity among two clustering results. ARI can yield low scores when a clustering result contains numerous small, yet coherent clusters. This does not necessarily indicate poor clustering quality, especially when the clusters are semantically meaningful. PUR assigns each cluster to the class that is most frequent in the cluster, measuring the accuracy of this assignment by counting the relative number of correctly assigned elements.

5 Evaluation results

In our evaluation, we used `Llama2Dictionary` and `Llama3Dictionary` with the parameters reported in Table 11 and `Flan-T5`. See Table 14 for specific parameters for each task.

5.1 Definition Generation (DG)

For the *Seen* benchmark evaluation, we consider the average performance over **WordNet** and **Oxford** (see Table 6). Note that, for **Wikitionary**, we do not compare with `Flan-T5` as the entire benchmark (i.e., Train-Dev-Test) has been used for training. Further details and comparisons with state-of-the-art methods across multiple benchmarks are reported in Table 15.

For `Flan-T5`, we report the original score presented by Giulianelli et al. (2023) (reported) and the score we obtain in our evaluation (observed). We believe that slight differences, where the observed results consistently under-perform compared to the reported results, are likely due to different parameter setting (e.g., temperature or greedy decoding). Nonetheless, the results are very similar.

Compared to `Flan-T5` observed, `LlamaDictionary` obtains higher results in all considered metrics. In addition, for reported, we achieve higher results for all metrics except BERT-F1, where our result is comparable (0.889 compared to 0.909). This is a interesting result considering that `Flan-T5` has been fine-tuned on more data than `LlamaDictionary`, i.e., all Train-Dev-Test sets of **Wikitionary**.

For the *Unseen* benchmarks, previous works have typically also used the data during training and are thus not fairly comparable. We report these results in Table 11. Thus we can evaluate only `Llama2Dictionary` and `Llama3Dictionary` and find that the latter consistently outperforms the former, unlike for the *Seen* benchmarks where the models were more even. This can be attributed to the fact that the Llama3-based model is larger than Llama2 in terms of parameters and training data.

For the *Unseen* benchmarks, the BERT-F1 scores, that rely on semantic similarity, are comparable to the *Seen* benchmarks. For the remaining scores, that rely on lexical overlap, the results for the *Unseen* benchmark is consistently, and significantly lower. We believe that this drop stems both from the issues discussed in Table 2 as well as the fact that the base Llama chat models, which have undergone *safety tuning*, are likely restricted from generating foul language, malicious, and toxic content that can be found in the Urban dictionary. Compared to the *Seen* benchmarks, the *Unseen* benchmarks also contain multi-word phrases for which the models have not been trained.

5.2 Word-in-Context (WiC)

Our results are reported in Table 7. Result using different SBERT models are summarized in Figure 2. Notably, we achieve a new state-of-the-art performance of .731 for the WiC task leveraging the definitions generated by `Flan-T5` + SBERT. The result by Bevilacqua et al. (2020) is particularly interesting for comparison, as it has also been obtained by relying on generated definitions.

	WordNet - **Oxford** *Seen*		**Urban** - **Wikipedia** *Unseen*	
	Llama2Dict.	Flan-T5 rep.	Llama2Dict.	-
	Llama3Dict.	Flan-T5 obs.	Llama3Dict.	Flan-T5 obs.
ROUGE-L	**.481**	.454	.161	-
	.400	.364	**.184**	.173
BLEU	**.402**	.257	.089	-
	.283	.266	**.100**	.095
BERT-F1	.880	**.909**	.764	-
	.889	.885	**.849**	**.849**
NIST	.938	-	.346	-
	.956	.828	**.405**	.339
SACREBLEU	**22.356**	-	4.823	-
	21.975	18.851	**5.484**	5.186
METEOR	.370	-	.151	-
	.426	.333	**.184**	.165
EX. MATCH	**50.161**	-	**.000**	-
	50.093	.110	**.000**	.000

Table 6: Average results for the **Definition Generation** task. The best results are highlighted in **bold**.

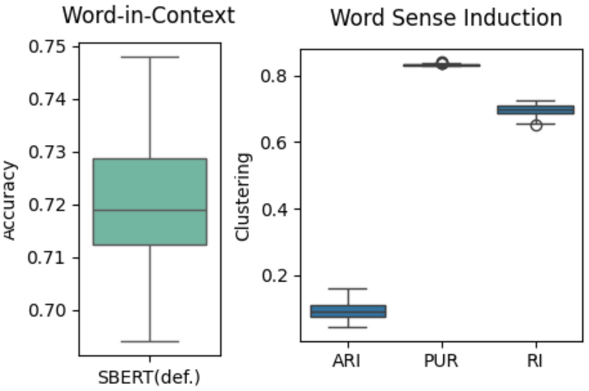

Figure 2: **Left**: Accuracy distribution on the base WiC task, using thirteen SBERT models. **Right**: ARI, PUR, and RI distribution on the WSI task, by considering our settings for the LSC task.

However, unlike our approach, they use multiple definitions per word usage. In contrast, we use a single definition per word usage, achieving higher results by employing both `LlamaDictionary` and `Flan-T5`.

As the WiC task requires distinguishing underlying meaning of word occurrences, the high performance of both `Flan-T5` and `LlamaDictionary` indicates that the use of definitions is a reasonable approach to capturing the intended sense while offering interpretability.

WiC	**Accuracy**
Levine et al. (2020)	.721
Bevilacqua et al. (2020)	.711
Peters et al. (2019)	.709
Chang and Chen (2019)	.692
`Flan-T5` + SBERT	**.731**
`Llama2Dictionary` + SBERT	.729
`Llama3Dictionary` + SBERT	.705

Table 7: Evaluation results for the **Word-in-Context** task. The best result is highlighted in **bold**.

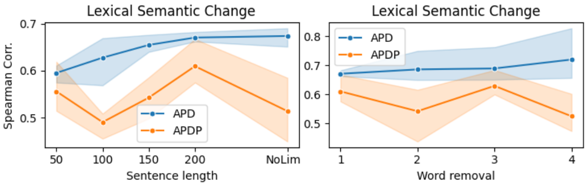

Figure 3: Avg. Spearman correlation by addressing LSC on different settings: different sentence length (**left**) and short word removal (**rigth**).

5.3 Lexical Semantic Change (LSC)

During our evaluation, we noticed that some of the annotated sentences present in the LSC benchmark were too long to be processed by our generative models (e.g., long word usages containing multiple sentences). This prompted us to evaluate the results by considering different sentence lengths, specifically 50, 100, 150 and 200 characters as well as the full sentences length. Our results are reported in Figure 3 and are consistently statistically significant. However, since we needed to discard up to 30% of sentences for `LlamaDictionary`, we proceeded with our experiments using up to 200 characters from each sentence.

Recent findings show that form-based approaches typically outperform sense-based approaches for the LSC task (Periti et al., 2024a) and that training models on WiC tasks enhances the modeling of lexical semantics (Arefyev et al., 2021). Similarly, we obtain higher performance for the form-based approach (APD, i.e., .662 – .682) than the sense-based one (APDP, i.e., .575 – .667), see Table 8. Although our results are lower than the established WiC-trained baselines, they are, on average, higher than those obtained using pretrained models (see Periti and Montanelli (2024) for an extensive overview). Additionally, we also note that processing the generated definitions by removing short words with fewer than 2, 3 or 4 characters, in addition to punctuation, consistently boosts the performance of `Flan-T5`, reaching correlations of .755, .762 and .827, respectively (see Figure 3). However, we did not observe the same boost for definitions generated by `LlamaDictionary`. After reviewing a small set of generated definitions, we hypothesize that this is due to the length of definitions generated by the models, with `LlamaDictionary` trained to provide *concise* definitions (See Figure 1).

When compared to state-of-the-art form-based approaches, our approach achieves medium-strong correlation results but does not outperform the considered baselines. When we consider APDP, the `Llama2Dictionary` model obtains the highest result, achieving a new state-of-the-art of .667 for interpretable LSC. This aligns with Giulianelli et al. (2023), who observe that the clusters of definitions have a lower intra-cluster dispersion compared to clusters using token and sentence embeddings.

LSC	method	Spearman
WiC-trained Aida and Bollegala (2024)	form-based	.774
WiC-trained Periti and Tahmasebi (2024)	form-based	**.886**
Keidar et al. (2022)	form-based	.489
Giulianelli et al. (2022)	form-based	.514
`Flan-T5` + SBERT	form-based	.682
`Llama2Dictionary` + SBERT	form-based	.667
`Llama3Dictionary` + SBERT	form-based	.662
WiC-trained Periti and Tahmasebi (2024)	sense-based	.652
Rother et al. (2020)	sense-based	.512
Montariol et al. (2021)	sense-based	.456
`Flan-T5` + SBERT	sense-based	.575
`Llama2Dictionary` + SBERT	sense-based	**.667**
`Llama3Dictionary` + SBERT	sense-based	.587

Table 8: Evaluation results for the **Lexical Semantic Change** task. The best result is highlighted in **bold**. Results are reported using both form-based and sense-based methods.

5.4 Word Sense Induction (WSI)

Our WSI evaluation relies on a recently developed benchmark originally designed for LSC. This benchmark contains cluster labels derived from manually annotated judgments of words *in-context*. These can therefore be considered as *silver* label data, rather than *gold* label data, as the clusters themselves have not been manually labeled.

Our results are reported in Table 9. We observe the highest results for the WiC-trained XL-LEXEME model (Cassotti et al., 2023), and GPT-4, were the training data is unknown and thus could include both WiC data and the WSI data used in this evaluation (Balloccu et al., 2024). When compared to standard pretrained models (i.e., BERT, mBERT, XLM-R), our results are consistently higher.

In line with Periti and Tahmasebi (2024), we observe low results in terms of ARI. We believe this stems from the quality of the original clusters to which we are comparing. The more flexible RI metric in Table 9 shows results comparable to the PUR scores.

In terms of the resulting clusters, we obtain an average number of clusters of 3.91 compared to the 9.61 of the original benchmark. This is in line with our intuition that definitions can be considered as prototypes of multiple word usages.

	model	ARI	PUR	RI
Results from Periti and Tahmasebi (2024)	BERT	.136	.700	.629
	mBERT	.067	.644	.526
	XLM-R	.068	.737	.582
	XL-LEXEME	.273	.834	.757
	GPT-4	**.340**	**.877**	**.802**
	`FlanT5`	.088	.832	.713
	`Llama2Dictionary`	.144	.835	.702
	`Llama3Dictionary`	.073	.832	.699

Table 9: Evaluation results for the **Word Sense Induction** task. The best result is highlighted in **bold**.

6 Conclusion

Inspired by recent advancements in text generation, in this work, we investigated the potential of fine-tuned large language models to generate sense definitions for words *in-context*. Specifically, we fine-tuned two new Llama chat based models, called `LlamaDictionary`, and assessed their performance along with an existing Flan-T5 model on the Definition Generation task. Next, we explored their utility for modeling word meaning by addressing lexical semantic tasks such as Word-In-Context, Word Sense Induction, and Lexical Semantic Change. In our experiments, we considered the generated definitions as intermediate representations, passed through a sentence embedding model.

Our results consistently show that we can use generated definitions to explicitly model the meaning of word usages through interpretable definitions. In all tasks, the use of sentence embeddings for generated definitions outperformed the use of standard token embeddings for word occurrences, setting new state-of-the-art results. Across tasks, we find that the use of the larger 7B and 8B `LlamaDictionary` models compared to the smaller 3B T5-based model obtain slightly higher results in the Definition Generation task, while being equally strong on the lexical semantics tasks. An extension of the `LlamaDictionary` models is to fine-tune them on all the benchmarks that have been used for the `Flan-T5` model, as well to fine-tune the models further on generated usage sentences (Malkin et al., 2021; Ma et al., 2024).

Our evaluation using automatically generated sense definitions in this paper paves the way for future advancements in modeling lexical semantics. For example, by offering an automatic labeling of senses, we can support the creation of lexicographic resources for all languages, including low-resource languages (Kong et al., 2022), providing a way to better know *what* change our words have experienced over time.

Limitations

In our work, we consider only English data as there are few available benchmarks, neither for training nor comparison on other languages. Given the necessary resources, we believe our approach to be language-agnostic and readily applicable to other languages.

We limited our experiments to `LlamaDictionary` and `Flan-T5` due to the cost and required computational resources for fine-tuning other large language models. We indeed exceeded the allocated resources on our National Super-computing during our experiments. Such large-scale models and experimental data must be approached cautiously as they will otherwise generate enormous computational costs (both in terms of monetary and environmental costs).

A further limitation of our models arises from the fact that existing Definition Generation benchmarks occasionally include multiple definitions for the same word meanings (e.g., Table 13). While this may serve as a form of regularization for training models, we believe that it may have influenced the uniformity in style and wording of our models. Unfortunately, statistics for these issues are non-existent. We thus advocate for further refinement to ensure consistency and coherence across definitions. We believe that, ideally, maximizing uniformity in definitions is desirable to develop models that offer consistent responses for similar word usages. This will be beneficial for any large-scale follow-up analysis relying on our evaluated approach.

In this paper, we integrated generated definitions with sentence embeddings. However, generated definitions often display higher lexical similarity to one another compared to word usages. Given the anisotropic nature of embedding spaces in large language models (Ethayarajh, 2019), the use of sentence embeddings might complicate discerning differences in definition of different complexity for language learners (Yuan et al., 2022). We thus believe future research should also explore the utilization of definition generation models alongside more conventional text-mining methods, such as count-based models. Count-based models may offer a more straightforward approach to processing interpretable, lexical similar definitions.

References

Eneko Agirre and Aitor Soroa. 2007. SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague, Czech Republic. Association for Computational Linguistics.

Taichi Aida and Danushka Bollegala. 2024. A Semantic Distance Metric Learning approach for Lexical Semantic Change Detection. *Preprint*, arXiv:2403.00226.

Domagoj Alagic, Jan Snajder, and Sebastian Pado. 2018. Leveraging Lexical Substitutes for Unsupervised Word Sense Induction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Marianna Apidianaki. 2023. From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation. *Computational Linguistics*, 49(2):465–523.

Nikolay Arefyev, Maksim Fedoseev, Vitaly Protastov, Daniil Homiskiy, Adis Davletov, and Alexander Panchenko. 2021. DeepMistake: Which Senses are Hard to Distinguish for a Word-in-Context Model. In *Proceedings of the Conference on Computational Linguistics and Intellectual Technologies (Dialogue)*, (online). RSUH.

Nikolay Arefyev and Vasily Zhikov. 2020. BOS at SemEval-2020 Task 1: Word Sense Induction via Lexical Substitution for Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 171–179, Barcelona (online). International Committee for Computational Linguistics.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating Scientific Definitions with Controllable Complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian’s, Malta. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Diego Bear and Paul Cook. 2021. Cross-Lingual Wolastoqey-English Definition Modelling. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 138–146, Held Online. INCOMA Ltd.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "How We Went beyond Word Sense Inventories and Learned to Gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

Leonard Bloomfield. 1933. *Language*. Holt, Rinehart and Winston, New York.

Dallas Card. 2023. Substitution-based Semantic Change Detection using Contextual Embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 590–602, Toronto, Canada. Association for Computational Linguistics.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Ting-Yun Chang and Yun-Nung Chen. 2019. What does this word mean? explaining contextualized embeddings with natural language definition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6064–6070, Hong Kong, China. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Are Synonym Substitution Attacks Really Synonym Substitution Attacks? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1853–1878, Toronto, Canada. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *Preprint*, arXiv:2305.14314.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using n-gram Co-occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kawin Ethayarajh. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional Generators of Words Definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271, Melbourne, Australia. Association for Computational Linguistics.

Noah Gardner, Hafiz Khan, and Chih-Cheng Hung. 2022. Definition Modeling: Literature Review and Dataset Analysis. *Applied Computing and Intelligence*, 2(1):83–98.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Mario Giulianelli, Andrey Kutuzov, and Lidia Pivovarova. 2022. Do Not Fire the Linguist: Grammatical Profiles Help Language Models Detect Semantic Change. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 54–67, Dublin, Ireland. Association for Computational Linguistics.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Maarten Grootendorst. 2022. BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure. *Preprint*, arXiv:2203.05794.

Muhammad Usman Hadi, Qasem al Tashi, Rizwan Qureshi, Abbas Shah, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, and Mubarak Shah. 2023. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Preprint*, arXiv:2403.14608.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *Preprint*, arXiv:2106.09685.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition Modelling for Appropriate Specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lawrence Hubert and Phipps Arabie. 1985. Comparing Partitions. *Journal of classification*, 2:193–218.

Mathew Huerta-Enochian. 2024. Instruction fine-tuning: Does prompt loss matter? *Preprint*, arXiv:2401.13586.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to Describe Unknown Phrases with Local and Global Contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3467–3476, Minneapolis, Minnesota. Association for Computational Linguistics.

John S. Justeson and Slava M. Katz. 1991. Co-occurrences of Antonymous Adjectives and Their Contexts. *Computational Linguistics*, 17(1):1–20.

Nikitas Karanikolas, Eirini Manga, Nikoletta Samaridi, Eleni Tousidou, and Michael Vassilakopoulos. 2024. Large Language Models versus Natural Language Understanding and Generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, PCI '23, page 278–290, , Lamia, Greece,. Association for Computing Machinery.

Kseniia Kashleva, Alexander Shein, Elizaveta Tukhtina, and Svetlana Vydrina. 2022. HSE at LSCDiscovery in Spanish: Clustering and Profiling for Lexical Semantic Change Discovery. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 193–197, Dublin, Ireland. Association for Computational Linguistics.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A survey of reinforcement learning from human feedback. *Preprint*, arXiv:2312.14925.

Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. Slangvolution: A Causal Analysis of Semantic Change and Frequency Dynamics in Slang. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1422–1442, Dublin, Ireland. Association for Computational Linguistics.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. ChatGPT: Jack of All Trades, Master of None. *Information Fusion*, 99:101861.

Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022. Multitasking Framework for Unsupervised Simple Definition Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5934–5943, Dublin, Ireland. Association for Computational Linguistics.

Matej Kosec, Sheng Fu, and Mario Michael Krell. 2021. Packing: Towards 2x NLP BERT Acceleration.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. Contextualized Embeddings for Semantic Change Detection: Lessons Learned. In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

Mike Lewis, 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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large Language Models Understand and Can be Enhanced by Emotional Stimuli. *Preprint*, arXiv:2307.11760.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xianghe Ma, Michael Strube, and Wei Zhao. 2024. Graph-based Clustering for Detecting Semantic Change Across Time and Languages. In *Proceedings of the 18th Conference of the European Chapter of* *the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1542–1561, St. Julian's, Malta. Association for Computational Linguistics.

Nikolay Malkin, Sameera Lanka, Pranav Goel, Sudha Rao, and Nebojsa Jojic. 2021. GPT Perdetry Test: Generating new meanings for new words. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5542–5553.

Christopher D Manning. 2009. *An Introduction to Information Retrieval*. Cambridge university press.

Leland McInnes, John Healy, and Steve Astels. 2017. HDBSCAN: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Timothee Mickus, Denis Paperno, and Matthieu Constant. 2019. Mark my Word: A Sequence-to-Sequence Approach to Definition Modeling. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 1–11, Turku, Finland. Linköping University Electronic Press.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Preprint*, arXiv:1301.3781.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and Interpretable Semantic Change Detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Juan Pablo Munoz, Jinjie Yuan, Yi Zheng, and Nilesh Jain. 2024. LoNAS: Elastic Low-Rank Adapters for Efficient Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10760–10776, Torino, Italia. ELRA and ICCL.

Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. 2021. A Comprehensive Survey on Word Representation Models: From Classical to State-of-the-Art Word Representation Language Models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(5).

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition Modeling: Learning to Define Word Embeddings in Natural Language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024a. Analyzing Semantic Change through Lexical Replacements. *Preprint*, arXiv:2404.18570.

Francesco Periti, Haim Dubossarsky, and Nina Tahmasebi. 2024b. (Chat)GPT v BERT Dawn of Justice for Semantic Change Detection. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 420–436, St. Julian's, Malta. Association for Computational Linguistics.

Francesco Periti and Stefano Montanelli. 2024. Lexical Semantic Change through Large Language Models: a Survey. *ACM Comput. Surv.* Just Accepted.

Francesco Periti, Sergio Picascia, Stefano Montanelli, Alfio Ferrara, and Nina Tahmasebi. 2023. Studying Word Meaning Evolution through Incremental Semantic Shift Detection: A Case Study of Italian Parliamentary Speeches.

Francesco Periti and Nina Tahmasebi. 2024. A Systematic Comparison of Contextualized Word Embeddings for Lexical Semantic Change. *Preprint*, arXiv:2402.12011.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge Enhanced Contextual Word Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Erika Petersen and Christopher Potts. 2023. Lexical Semantics with Large Language Models: A Case Study of English "break". In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 490–511, Dubrovnik, Croatia. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2021. *Contextualized Embeddings*, pages 69–96. Springer International Publishing, Cham.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

James Pustejovsky and Branimir Boguraev. 1993. Lexical Knowledge Representation and Natural Language Processing. *Artificial Intelligence*, 63(1):193–223.

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

William M. Rand. 1971. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

David Rother, Thomas Haider, and Steffen Eger. 2020. CMCE at SemEval-2020 Task 1: Clustering on Manifolds of Contextualized Embeddings to Detect Historical Meaning Shifts. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 187–193, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Vincent Segonne and Timothee Mickus. 2023. Definition Modeling : To model definitions. Generating Definitions With Little to No Semantics. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 258–266, Nancy, France. Association for Computational Linguistics.

Lutfi Kerem Senel, Ihsan Utlu, Veysel Yucesoy, Aykut Koc, and Tolga Cukur. 2018. Semantic Structure and Interpretability of Word Embeddings. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 26(10):1769–1779.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arXiv:2307.09288.

Yu-Hsiang Tseng, Mao-Chang Ku, Wei-Ling Chen, Yu-Lin Chang, and Shu-Kai Hsieh. 2023. Vec2Gloss: definition modeling leveraging contextualized vectors with Wordnet gloss. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 679–690.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International*

Conference on Neural Information Processing Systems, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Hendryk Weiland, Maike Behrendt, and Stefan Harmeling. 2023. Automatic Dictionary Generation: Could Brothers Grimm Create a Dictionary with BERT? In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 102–120, Ingolstadt, Germany. Association for Computational Lingustics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jiaxin Yuan, Cunliang Kong, Chenhui Xie, Liner Yang, and Erhong Yang. 2022. COMPILING: A Benchmark Dataset for Chinese Complexity Controllable Definition Generation. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 921–931, Nanchang, China. Chinese Information Processing Society of China.

Hengyuan Zhang, Dawei Li, Yanran Li, Chenming Shang, Chufan Shi, and Yong Jiang. 2023. Assisting Language Learners: Automated Trans-Lingual Definition Generation via Contrastive Prompt Learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 260–274, Toronto, Canada. Association for Computational Linguistics.

Hengyuan Zhang, Dawei Li, Shiping Yang, and Yanran Li. 2022. Fine-grained Contrastive Learning for Definition Generation. In *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 (Volume 1: Long Papers)*, pages 1001–1012, Online only. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction Tuning for Large Language Models: A Survey. *Preprint*, arXiv:2308.10792.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Fine-tuning

In our experiments, we conducted multiple rounds of fine-tuning, systematically testing various parameters. Specifically, we detail these configurations in Table 10. In line with Huerta-Enochian (2024), who recently demonstrated that prompt loss can be safely ignored for many datasets, we observed lower preliminary results in the evaluation tasks for models chosen based on validation performance. Therefore, we selected the final models (see Table 11) based on the checkpoint from the last training epoch that had the best performance on the Definition Generation task.

Parameter	Experimented values
Model	`meta-llama/Meta-Llama-3-8B-Instruct`, `meta-llama/Llama-2-7b-chat-hf`
GPU	A100:fat (80 GB)
Hours	7-8
PEFT	LoRA, QLoRA
Dropout	0.05, 0.1, 0.2
Weight decay	0.001, 0.0001
Learning rate	1e-4, 1e-5
Lora ranks	8, 32, 64, 128, 256, 512, 1024
Lora alpha	16, 64, 256, 512, 1024, 2048
Warmup ratio	0.03, 0.05
Eval steps	250
Train epochs	4, 5, 10
Max seq. length	512
Batch size	32
Optimizer	Adam
LoRA target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj, lm_head

Table 10: Settings and parameters used during training. Parameters shown in small font represent preliminary experiments that were not further evaluated.

Final setting	Llama2Dictionary	Llama3Dictionary
GPU	A100:fat (80 GB)	A100:fat (80 GB)
Hours	7-8	8-9
PEFT	LoRA	LoRA
Dropout	0.1	0.05
Weight decay	0.001	0.001
Learning rate	1e-4	1e-4
Lora ranks	1024	512
Lora alpha	2048	1024
Warmup ratio	0.05	0.05
Eval steps	epochs	epochs
Train epochs	4	4
Max seq. length	512	512
Batch size	32	32
Optimizer	Adam	Adam
LoRA target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj, lm_head	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj, lm_head

Table 11: Parameters of our final models. Our code will be publicly available for further details. For finetuning, we rely on the transformers library (Wolf et al., 2020).

A.1 Lora rank-alpha

We conduct fine-tuning using LoRA, (Hu et al., 2021) and QLORA, (Dettmers et al., 2023) obtaining very similar evaluation results. Drawing from insights from prior research (Munoz et al., 2024) as well recent online discussions, we adopted a strategy where the LoRA alpha α was set to double the LoRA rank r. In our experiments for the Definition Generation task, larger ranks resulted in higher performance on **WordNet** and slightly higher performance on **Oxford** benchmarks. However, no improvement was noted for **Wiktionary** (see Figure 4).

B SBERT models

In our experiments, we made an effort to evaluate all the Bi-Encoder SBERT models available at `https://sbert.net/` (see Table 12). This thorough assessment ensures that our findings are robust and accurate. While we acknowledge that other models may exist, the evaluation results we present remain valuable and consistent across the models tested, contributing to the broader perspective presented in the paper.

Further parameters are related to our procedure for addressing the Word-in-Context, Word Sense Induction, and Lexical Semantic Change tasks. We report these parameters in Table 14.

SBERT models
`all-mpnet-base-v2`
`multi-qa-mpnet-base-dot-v1`
`all-distilroberta-v1`
`all-MiniLM-L12-v2`
`multi-qa-distilbert-cos-v1`
`all-MiniLM-L6-v2`
`multi-qa-MiniLM-L6-cos-v1`
`paraphrase-multilingual-mpnet-base-v2`
`paraphrase-albert-small-v2`
`paraphrase-multilingual-MiniLM-L12-v2`
`paraphrase-MiniLM-L3-v2`
`distiluse-base-multilingual-cased-v1`
`distiluse-base-multilingual-cased-v2`

Table 12: Experimented SBERT models. We report in **bold** the model used for the results obtained in the main paper. We use this model as it was used in previous experiments by Giulianelli et al. (2023).

C Definition Generation

In our work, we extensively evaluated our `LlamaDictionary` models along with the `Flan-T5` models by Giulianelli et al. (2023), setting new state-of-the-art results on the Definition Generation tasks across multiple benchmarks. In Table 15, we provide a full comparison, including individual scores for each benchmark and the measures considered.

Benchmark	**Target** w	**Example** e	**Definition** e
WordNet	accuracy	He was beginning to doubt the *accuracy* of his compass	The quality of being near to the true value
Oxford	accuracy	However, these studies have not generally had enough participants to provide precise estimates of *accuracy*.	The quality or state of being correct or precise
Wiktionary	accuracy	The efficiency of the instrument will also depend upon the *accuracy* with which the piston fits the bottom and sides of the barrel. When the piston is depressed to the bottom, it is considered in theory to be in absolute contact, so as to exclude every particle of air from the space between it and the bottom.	The state of being accurate; being free from mistakes, this exemption arising from carefulness; exactness; correctness
Oxford	yesterday	*Yesterday* the weather was beautiful	On the day preceding today
Oxford	yesterday	It was in *yesterday* 's newspapers	The day immediately before today
Oxford	yesterday	I am doing a research paper on women 's voting rights ; *yesterday* and today	On the day before today
Oxford	yesterday	On a day like today after *yesterday* , i tend to reflect , internalize , and re-address the balance	The day before today

Table 13: Example of correct but inconsistent definitions from the considered benchmarks. It is unnecessary to train the model to provide different answers. Ideally, a single definition should be used for different examples of the considered target.

	Evaluation tasks			
	DG	**WiC**	**WSI**	**LSC**
gen. model	LlamaDictionary, Flan-T5	LlamaDictionary, Flan-T5	LlamaDictionary, Flan-T5	LlamaDictionary, Flan-T5
temperature	0.0	0.0	0.0	0.0
enc. model	roberta-large	all-distilroberta-v1	all-distilroberta-v1	all-distilroberta-v1
metric	BERTScore	cosine	cosine	cosine (APD) canberra (APDP) following Periti et al.; Periti and Tahmasebi
clustering	-	-	HDBSCAN	HDBSCAN
HDBSCAN-allow_single_cluster	-	-	True	True
HDBSCAN-min_cluster_size	-	-	2	2
HDBSCAN-cluster_selection_method	-	-	leaf	leaf

Table 14: Models and parameters used for addressing the DG, WIC, WSI, and LSC tasks. We rely on the HDBSCAN implementation of the scikit-learn library (Pedregosa et al., 2011).

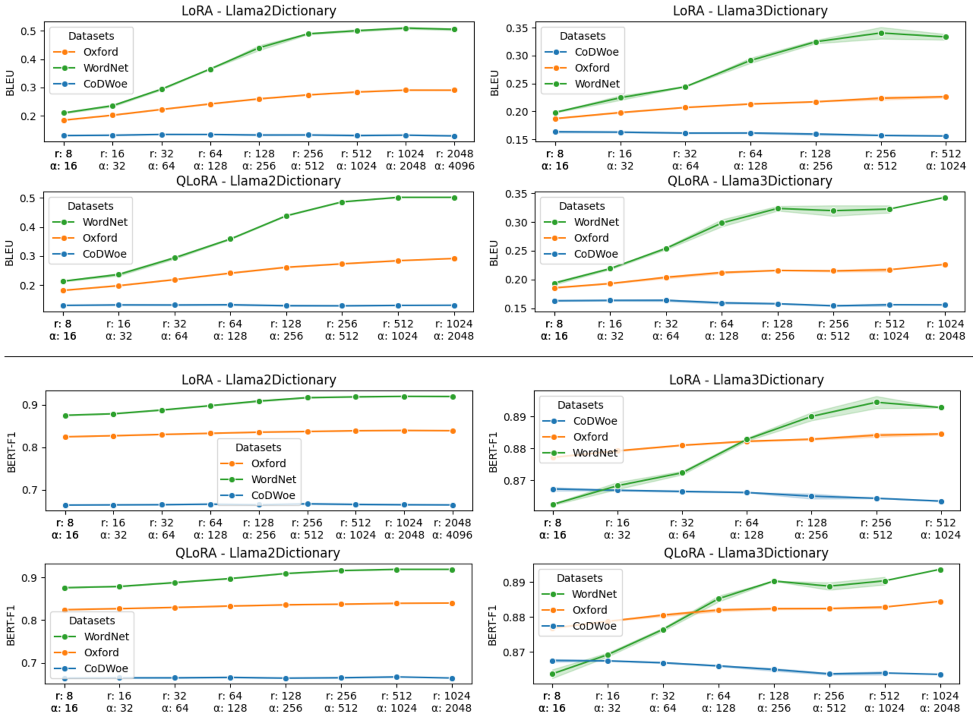

Figure 4: Average performance of trained models using LoRA (Hu et al., 2021) and QLoRA (Dettmers et al., 2023) with parameters from Table 10. We conducted experiments with LoRA *alpha* α set to double the *rank* r and observed that larger ranks resulted in higher performance on **WordNet** and slightly higher performance on **Oxford** benchmarks. However, no improvement was noted for **Wiktionary**. We report BERT-F1 and BLEU as examples. Similar trends were observed for other performance metrics.

	ROUGE-L	BLEU	BERT-F1	NIST	SACREBLEU	METEOR	EXACT MATCH
WordNet - seen							
Noraset et al. (2017)	-	.236*	-	.497*	-	-	-
Ni and Wang (2017)	-	.248*	-	.403*	-	-	-
Gadetsky et al. (2018)	-	.237*	-	.443*	-	-	-
Ishiwatari et al. (2019)	-	.248	-	.435*	-	-	-
Huang et al. (2021)	-	.327	-	.646	-	-	-
Zhang et al. (2022)	-	.320	-	.747	-	-	-
Giulianelli et al. (2023) Reported	.522	.328	**.921**	-	-	-	-
Giulianelli et al. (2023) Observed	.405	.320	.893	.907	23.302	.374	.164
Llama2chat	**.564**	**.513**	**.920**	**1.391**	**41.096**	**.536**	**.373**
Llama3Instruct	.435	.339	.893	1.012	27.400	.480	.131
Oxford - seen							
Noraset et al. (2017)	-	.149*	-	.327*	-	-	-
Ni and Wang (2017)	-	.176*	-	.313*	-	-	-
Gadetsky et al. (2018)	-	.120	-	.358*	-	-	-
Ishiwatari et al. (2019)	-	.185	-	.382*	-	-	-
Huang et al. (2021)	-	.265	-	.742	-	-	-
Bevilacqua et al. (2020)	.294	.088	.768	-	-	.135	-
Zhang et al. (2022)	-	.271	-	.794	-	-	-
Giulianelli et al. (2023) Reported	.387	.186	**.897**	-	-	-	-
Giulianelli et al. (2023) Observed	.324	.213	.878	.749	14.400	.292	.057
Llama2chat	**.398**	**.291**	.840	**.969**	**21.410**	.367	**.158**
Llama3Instruct	.365	.228	**.885**	.900	16.550	**.373**	.055
Wikitionary - seen							
Llama2chat	.222	.131	.666	.408	6.963	.183	.025
Llama3Instruct	**.267**	.156	**.863**	.517	**8.100**	**.232**	**.034**
Urban - unseen							
Noraset et al. (2017) - seen	-	.515*	-	.104*	-	-	-
Ni and Wang (2017) - seen	-	**.899***	-	.174*	-	-	-
Gadetsky et al. (2018) - seen	-	.088*	-	.194*	-	-	-
Ishiwatari et al. (2019) - seen	-	.105	-	.192*	-	-	-
Huang et al. (2021) - seen	-	.177	-	.355	-	-	-
Zhang et al. (2022) - seen	-	.194	-	**.410**	-	-	-
Giulianelli et al. (2023) - unseen Observed	.106	.053	.835	.167	2.160	.068	**.001**
Llama2chat - unseen	.110	.055	.812	.170	**2.247**	.071	**.001**
Llama3instruct - unseen	**.115**	.057	**.836**	.197	2.331	**.079**	**.001**
Wikipedia - unseen							
Noraset et al. (2017) - seen	-	.446*	-	.334*	-	-	-
Ni and Wang (2017) - seen	-	.527*	-	.552*	-	-	-
Gadetsky et al. (2018)- seen	-	.450*	-	.331*	-	-	-
Ishiwatari et al. (2019)- seen	-	.538	-	.567*	-	-	-
Huang et al. (2021)- seen	-	**.556**	-	**.640**	-	-	-
Giulianelli et al. (2023) - unseen Observed	.240	.138	**.863**	.511	8.212	.263	**.000**
Llama2chat - unseen	.213	.123	.716	.523	7.399	.232	**.000**
Llama3Instruct - unseen	**.253**	**.144**	**.863**	.614	**8.638**	**.290**	**.000**

Table 15: Evaluation results for the **Definition Generation** task. The best result is highlighted in bold. Our model is trained exclusively on the training set of the WordNet, Oxford, and Wiktionary datasets. Results marked with * are reported from experiments in Huang et al. (2021).