

Preference Optimization Improves Definition Generation: Semantic Change in EU Institutional Sustainable Finance Disclosures

Mariia Fedorova
University of Oslo
mariiaf@ifi.uio.no

Andrey Kutuzov
University of Oslo
andreku@ifi.uio.no

Gustavo Cadenas Delascio
University of Oslo
g.c.delascio@globe.uio.no

Abstract

Dictionary-like definition generators are known to be useful for lexical semantic change detection (LSCD). However, they suffer from overly granular sense distribution, producing excessive number of definitions for one and the same lexical sense. In this work, we show that fine-tuning with Direct Preference Optimization (DPO) helps to overcome this problem. We test our approach both quantitatively on a curated academic LSCD benchmark and qualitatively on real-world data for a practical task of auditing diachronic changes in sustainability terminology within project documentation of EU development finance. All models, code and predictions produced in this study are publicly available at <https://github.com/lgtoslo/dpo-for-definition-generation>.

1 Introduction

Diachronic lexical semantic change is a change of the meanings of words and phrases throughout time. While Average Pairwise Distance (APD) on target word embeddings (Giulianelli et al., 2020) is currently the recommended method of change detection (Schlechtweg et al., 2025), generated dictionary-like definitions allow for clustering word usage examples by senses and labelling these word senses with human-readable explanations (Kutuzov et al., 2024). This paves the way for explainable and interpretable semantic change modelling, potentially useful in practice for researchers from fields other than NLP.

Definition generators are large language models which undergo supervised fine-tuning (SFT) on datasets of target words used in specific contexts and human-produced definitions of target words in these contexts (Giulianelli et al., 2023; Fedorova et al., 2024; Tatarinov and Demidovsky, 2025).¹

¹E.g., PROMPT: ‘He sat on the bank of the river and cried. What is the definition of bank?’, RESPONSE: ‘the rising ground bordering a lake, river, or sea or forming the edge of a cut or hollow’

However, a known problem of definition generators is producing slightly different definitions for the same sense of a target word used in different contexts: the definitions may differ by only one word or even by one punctuation sign, while their semantics is the same or very similar. Fedorova et al. (2024) give an example of the English word ‘plane’ being defined either as ‘An aircraft, especially one designed for military use’ or ‘An aircraft, especially a military aircraft’. These ‘noise senses’ make it difficult for domain experts to analyse the resulting predictions and to draw conclusions about the nature of specific change cases.

We approach this problem using preference optimization (Ouyang et al., 2022) of pre-trained definition generators. The general idea is that the desired behavior of the model is to generate very similar (or, ideally, identical) definitions for a target word used in the same sense. Thus, in our preference learning data, we choose one definition per sense as the preferred response and label other definitions (generated by the same model) as the rejected responses. Then we employ DPO (Rafailov et al., 2023) to push the model towards generating more uniform definitions per sense.

We find that fine-tuning definition generators with DPO does consistently reduce the number of definitions per target word (thus, the number of inferred senses), compared to the base models (which only underwent SFT). The picture is more nuanced for the performance on the graded semantic change benchmark used in SemEval’20 Task 1 (Schlechtweg et al., 2020). Here, DPO can either improve or degrade the performance, depending on the base model family (see below).

Additionally, we conduct qualitative analysis of the performance of the resulting models on a domain-specific dataset of EU documents. We used a pipeline combining definition generation and the state-of-the-art XL-Durel model (Yadav and Schlechtweg, 2025) to provide a domain ex-

pert a list of terms along with the dynamics of their senses changing over time.

Our contributions can be summarized as follows:

- We show that preference optimization for less diverse definitions per sense improves the sense granularity of definition generators.
- It also improves the performance of the definition generator on semantic change detection, unless the model is already strong.
- We train a new family of definition generators based on OLMo-3-7B model, outperforming prior work based on Llama3-8B.
- We qualitatively evaluate the resulting models on the case of sustainability analysis on the corpus of European Investment Bank (EIB) documents, with a domain expert confirming our findings.

2 Related work

Semantic change The task of diachronic semantic change detection (SCD) boils down to (given a set of n historical corpora) either binary classification of a set of target words to either ‘changed’ or ‘not changed’, or to ranking of the target words by the degree of their semantic change (graded semantic change detection). One of the first standard sets for evaluating SCD was established in SemEval 2020 Task 1 (Schlechtweg et al., 2020); we use its English subset to sanity-check that our definition generators do not perform worse on it than the ones from the previous work. Recently, contextualized definition generation was shown to be promising for modelling word meaning, and, subsequently, for tracing semantic change (Periti et al., 2025). This is the direction we pursue in this paper. However, the state-of-the-art approaches to semantic change use embedding vectors as sense representations. For example, XL-Durel is a word-in-context word sense disambiguation model, accepting a sentence with a specially tagged target word, e.g. She sat on the <t>bank </t> of the river to rest². The model’s output are token embeddings of target words which are used as the APD input.

Definition generation Giulianelli et al. (2023) and Cassotti et al. (2024) pioneered employing definition generation for semantic change detection, with T5 encoder-decoder models. Periti et al.

(2024) fine-tuned decoder-only Llama 3 model on the same English data, excluding the test split³. Periti et al. (2025) continued this line of work with English⁴ and multilingual models trained on the DBnary dataset, consisting of Wiktionary⁵ data (Sérasset, 2012). Fedorova et al. (2025) compared encoder-decoder and decoder-only models.

Sustainability analysis Recent work establishes language models as a standard method for auditing corporate sustainability disclosures. For example, Luccioni et al. (2020) utilise RoBERTa-based question-answering models to extract climate-risk passages from heterogeneous financial reports, guided by specific disclosure frameworks. Similarly, Schimanski et al. (2024) develop specialised language models to classify corporate communication into distinct environmental, social, and governance categories, demonstrating a correlation between disclosure volume and external ESG ratings.

3 Datasets for supervised fine-tuning and preference optimisation

We use the same training and development splits of DBnary⁶ as (Periti et al., 2025) and further divide them into SFT and DPO subsets (the SFT subset was only used for the OLMo-3 model). The splits were stratified by the number of unique gold senses. Their statistics are to be found in Table 1. It is also worth noting that DBnary includes definitions both of single words and multi-word entities (‘Leave behind’, ‘Stroke of work’ etc.).

An example of a SFT training instance is to be found in Appendix B.1.

The DPO subset is used by the SFT-tuned models to produce chosen-rejected sample pairs as follows: we generate model’s responses for all usage examples and post-process the resulting dataset so that ‘preferable’ responses are always the same for usage examples of the same sense. ‘Preferable’ definitions are chosen by the shortest Levenshtein distance (Levenshtein et al., 1966) to the gold one. E.g., for the target word ‘juvenocracy’, the gold sense definition is ‘Government by youth’. For

³<https://hf.co/FrancescoPeriti/Llama3Dictionary-merge>

⁴https://hf.co/FrancescoPeriti/LlamaDictionary-en_ML38BI

⁵https://en.wiktionary.org/wiki/Wiktionary:Main_Page

⁶<https://hf.co/datasets/FrancescoPeriti/Dbnary-en>. DBnary is licensed under Creative Commons Attribution-ShareAlike 3.0.

²<https://huggingface.co/sachinn1/xl-durel>

Split	Target words	Examples
SFT train	90 030	220 325
SFT dev	5 144	14 793
DPO train	90 030	220 540
DPO dev	5 145	14 407

Table 1: DBnary data split statistics before post-processing the DPO split.

three different usages of ‘juvenocracy’ in this sense, a definition generator produced the following three definitions: 1) ‘A government by youth or children’, 2) ‘Government by youth; government of children’, 3) ‘Government by youth; government by children or young people’. Since definition 1 is most similar to the gold sense definition by edit distance, it is labelled as preferred; the rest of the definitions (2 and 3) are labelled as rejected, thus forming two preference learning pairs (1 vs 2 and 1 vs 3). After this post-processing, only senses with at least two usage examples are kept, with the resulting size of 20 016 examples in the train split and 2 086 in the development split.

An example of a DPO training instance is to be found in Appendix B.2.

4 Domain-specific corpus: EIB Environmental and Social Data Sheets

In addition to evaluating our model on standard academic LSCD benchmarks, we also conducted qualitative analysis of their capabilities to trace sense changes in a domain-specific corpus. This section describes the corpus, for the analysis itself see Section 6.

The EIB is the EU’s biggest multilateral development bank and the main implementer of the Global Gateway strategy. Its lending decisions directly operationalise EU external investment policy at the project level. The Environmental and Social Data Sheet (ESDS) is a mandatory technical document produced for every EIB-financed project during the appraisal phase. It constitutes the environmental and social due diligence record submitted to the Board of Directors for financing approval (European Investment Bank, 2015).

Unlike press releases or public declarations, ESDS are operational instruments that influence capital allocation. Therefore, semantic changes within this corpus reflect actual bureaucratic implementation rather than political signalling. The

corpus spans from 2014 to 2026. The starting point is EU directive 2014/52/EU and the Aarhus Regulation EC N°1367/2006, which regulate the assessment of projects on the environment and prompted the EIB to establish a public register. This window establishes the historical baseline of EIB sustainability discourse before capturing the 2021 launch of the Global Gateway and the concurrent repositioning of the EIB as a “climate bank”. This allows us to distinguish semantic shifts associated with specific geopolitical mandates from broader baseline drift in sustainability vocabulary.

The ESDS collection is scraped from EIB website⁷ (4 201 documents). Since language tags on the website has turned out to be not reliable, we apply language identification on the document level using OpenLID-v3 model⁸ (Burchell et al., 2023; Fedorova et al., 2026). 4 189 documents are in English, with the rest of 12 documents being in Albanian, Greek, and French. Thus, our study has to be limited to English only. The documents are split into sentences by punctuation marks, resulting in 151 041 sentences. These are used as word usage examples for LSCD.

We apply flair/chunk-english⁹ (Akbik et al., 2018) model to extract all noun phrases in the corpus. Out of those, a domain expert samples target phrases of interest. This sample is used for qualitative analysis in Section 6.

5 Experiments

5.1 Baselines

For evaluation on SemEval’20 Task 1, we use the following baseline models, all of which underwent SFT on definition generation for English:

- mT0-xl (4 billion parameters) (Fedorova et al., 2024)
- Llama3Dictionary (8 billion parameters) (Periti et al., 2024)
- LlamaDictionary-en_ML38BI (8 billion parameters) (Periti et al., 2025)

⁷<https://www.eib.org/en/registers/all/index>. The website’s terms of use <https://www.eib.org/en/terms-of-use.htm> allow downloading data, but are not clear about republishing/creating derivatives for research purposes. Thus, we will only publish the code to reproduce our processing, but not the collection itself.

⁸<https://hf.co/HPLT/OpenLID-v3>

⁹<https://huggingface.co/flair/chunk-english>

The first two models were fine-tuned on a mixture of WordNet (Ishiwatari et al., 2019), Oxford (Gadetsky et al., 2018), and CoDWoE (Mickus et al., 2022) datasets, while the third one was fine-tuned on DBnary. All of them were successfully used in prior work on definition generation and semantic change modeling.

5.2 Our definition generators

We extend prior work by creating three new definition generators. First, we preference-optimize LlamaDictionary-en_ML38BI model. This was an obvious experiment in order to test whether DPO can improve a definition generator. However, one downside of Llama-based LLMs is the lack of full openness: the training data of Llama models is unknown, and the model license is restrictive. Thus, we also used a more recent and fully open language model of about the same size: Olmo-3-7B-Instruct-SFT¹⁰ (Olmo et al., 2025).

The original Olmo-3-7B-Instruct-SFT (7 billion parameters) was not specifically fine-tuned to generate contextualized definitions, so we first ran on it a round of SFT, using the corresponding DBnary split (see Section 3). After that, it was preference-optimized in the same way as LlamaDictionary-en_ML38BI.

Thus, the three models we introduce here are Llama-DPO, OLMo-SFT and OLMo-DPO.

5.3 Hyperparameters and hardware

We used HuggingFace trl¹¹ library for both SFT and DPO. All the DPO fine-tuning was done in one epoch using LoRA (Hu et al., 2022) and bfloat16 (bf16) precision to reduce the required compute. The SFT round on OLMo-3 was run on maximum 20 epochs (in practice, early-stopped after epoch 2). Table 2 presents the hyperparameters of both SFT and DPO.

Note that our LoRA alpha and rank parameters for DPO are much lower than for SFT, since at this step the model should change only slightly. We also found that too aggressive DPO (when using β in the range 0.4-0.1) increases predictions’ diversity rather than declines it (we track the average number of sense per word on the development data split during training), thus we use the value 0.5.

DPO was done on a single NVIDIA GH200 GPU for Llama3-based models, and on two GPUs for

Parameter	SFT	DPO
LoRA rank	256	64
LoRA alpha	512	16
LoRA dropout	0.1	0.2
learning rate schedule	linear	linear
learning rate	2e-5	1e-6
batch size	32	24

Table 2: Training hyperparameters.

OLMo-3 based ones¹². 1 epoch of DPO took not more than 2 hours. SFT was done on two NVIDIA GH200 GPUs, with 5 hours walltime.

5.4 Change detection

The definition generators described in the previous sections were used to produce contextualized definitions for selected occurrences of target words in SemeEval’2020 English corpora (historical periods 1810–1860 and 1960–2010). The generated definitions serve as interpretable representations of the target words’ semantics in context. We fully re-use the ‘definitions as senses’ change detection algorithm proposed in (Fedorova et al., 2024), employing Jensen-Shannon divergence (JSD) to compare sense distributions across time periods, and applying definition generators to exactly the same usage examples.

Our hypothesis is that preference-optimized definition generators will (at least somewhat) alleviate the issue of redundant definitions per sense. Fedorova et al. (2024) already tried to address this issue by merging together definitions with low Levenshtein edit distance between them. But they could not tune the merging threshold hyper-parameter because of the lack of development dataset. In this work, we use the DBnary development split for this purpose, so we tune merging threshold on it (optimizing the number of senses per target word), before final evaluation on the SemEval’2020 test set. The resulting optimal threshold with full merging was found to be 22, and we refer to it below as ‘tuned threshold’.

Since we expect that a surface metric like an edit distance would not capture cases where two definitions express the same meaning using different words, we also tried Agglomerative Clus-

¹⁰<https://hf.co/allenai/Olmo-3-7B-Instruct-SFT>

¹¹<https://huggingface.co/docs/trl>

¹²It is absolutely possible to fine-tune OLMo-3 on a single GPU as well.

tering (Kaufman and Rousseeuw, 2009) of definitions’ embeddings produced by BAAI/bge-m3 model (Chen et al., 2024)¹³ with the complete linkage and cosine distance threshold 0.2.

Finally, for the qualitative analysis in Section 6, we re-rank the target words according to their change score using APD on the embeddings obtained¹⁴ from the state-of-the-art word-in-context model XL-Durel (Yadav and Schlechtweg, 2025), so that the domain expert gets true positive changes on top of their list.¹⁵ The mixture of APD and ‘definitions-as-senses’ lets the user not only see which words have changed, but also which exact senses were gained and which ones were lost.

Predictions by BAAI/bge-m3 and XL-Durel run on a single Google Colab’s T4 GPU in not more than 15 minutes.

5.5 Results

Table 3 shows Semeval 2020 Task 1 subtask 2 (graded semantic change detection) results for baselines and our models. We only focus on subtask 2, because it better suits our research problem of finding change of non-discrete senses in the ESDS data. In this work, we report the results with no merging, with the merging parameter from (Fedorova et al., 2024) (threshold 50), with the optimal merging threshold tuned on DBnary development set (22, full merging), and with two strategies of clustering (done separately for each time period – we tried it because in the original paper it was shown to perform better – and on all available predictions for both time periods).

One should note that Semeval 2020 Task 1 is a small dataset: it only contains 37 target words. As Phan-Tat et al. (2026) point out, this limitation makes it impossible to fairly rank the performance of different systems outputting close scores, because these scores yield overlapping confidence intervals (CIs). Following this work, we have calculated CIs for the scores of our models. We find no consistent improvement from merging or clus-

tering across models, training data and fine-tuning approaches; the best absolute score is achieved without any postprocessing. The models fine-tuned on DBnary perform slightly better than those fine-tuned on the Wordnet-Oxford-CoDWoE mixture following (Giulianelli et al., 2023). However, these scores do not allow to conclude whether OLMo or LlamaDictionary-ML38B is a better base model choice, or whether the DPO step is necessary, since their scores are within the intersection of their CIs. Overall, definition generation does not outperform sentence embedding approaches such as XL-Durel+APD (either because of high granularity or because word usages representing some senses were left behind in our samples); but our models have passed the sanity check making them preferable over previous work.

Table 4 gives more details on the influence of DPO on definition generation models. We use DBnary development set (5 145 target words, 14 407 usage examples) as a human performance baseline to compare our models against. Specifically, we count the average number of senses (unique definitions) per target word, the ratio of ambiguous target words and the average number of senses per *ambiguous* target word.

Distribution of the senses in DBnary and models’ predictions differs a lot. First, we see that definition generators overestimate the number of senses per word and the ratio of ambiguous words (i.e., they yield more than one sense for some monosemantic words). Applying DPO reduces number of predicted senses, confirming our original hypothesis. The average number of senses per ambiguous word is mirroring this dynamics (if we are keeping the total number of senses fixed and increasing the share of ambiguous words, the average ambiguity of these words goes down). Thus, DPO does push the models towards the gold DBnary distribution.

The ability to reproduce sense distribution in one dataset may not directly translate to higher performance in semantic change detection on another dataset. OLMo-3 models are much more prone to generating diverse definitions than Llama models: the percentage of ambiguous words is more than twice higher than in gold DBnary distribution.

6 Case study: qualitative sustainability analysis

In sustainability studies, core terminology is highly contested and semantically ambiguous (Connelly,

¹³BAAI/bge-m3 is a sentence encoder for information retrieval. It accepts sentences ‘as is’, without any preprocessing, as an input.

¹⁴We used the lemmatizer based on HPLT 3.0 (Oepen et al., 2026) BERT model (https://hf.co/HPLT/hplt_gpt_bert_base_3_0_eng_Latn-UD) for finding the target words in the usage examples.

¹⁵This model was trained to produce sentence embeddings of usage examples with a target word surrounded by special symbols. We were able to reproduce XLD-APD performance from Figure 5 in (Schlechtweg et al., 2025), achieving the score of 0.768.

Model	mT0-xl SFT	Llama3Dictionary SFT	LlamaDictionary-ML38B SFT	OLMo-3 SFT	LlamaDictionary-ML38B DPO	OLMo-3 DPO
No merging	0.476 [0.18, 0.69]	0.429 [0.12, 0.66]	0.517 [0.23, 0.72]	0.586 [0.32, 0.76]	0.532 [0.25, 0.73]	0.546 [0.27, 0.74]
Minimalist merging 50	0.565	0.547	0.237	0.445	0.354	0.258
Full merging 50	0.493	0.494	0.200	0.389	0.224	0.170
Full merging 22 (tuned)	0.431	0.510	0.444	0.562	0.426	0.380
Clustering	0.435	0.465	0.464	0.557	0.493	0.498
Clustering separately	0.171	0.214	0.42	0.207	0.458	0.068

Table 3: Evaluation on SemEval’20 Task 1, Spearman rank correlations. Generated definitions were obtained in single runs, and sense detection algorithms are deterministic. Confidence intervals for the best performing postprocessing (no postprocessing) are reported in brackets.

Model	Total senses	Senses per word	% amb.	Senses per amb.
Human (DBnary gold)	7 645	1.49	20.1	6.76
Llama ML38B SFT	9 228	1.79	30.2	5.50
Llama ML38B DPO	9 044	1.76	29.3	5.58
OLMo-3 SFT	12 721	2.47	46.7	4.68
OLMo-3 DPO	12 143	2.36	45.3	4.72

Table 4: Distribution of senses in the DBnary development set and in the predictions of our models. ‘Amb.’ stands for ‘ambiguous words’ (more than one sense).

2007). Policy institutions selectively leverage these elastic definitions to classify investments and justify environmental action. By stretching definitions to satisfy conflicting pressures, institutions maintain strategic ambiguity (Leitch and Davenport, 2007). This ambiguity generates a significant divergence between an institution’s discursive commitments and its operational behaviour (Fiss and Zajac, 2006). Consequently, ambitious sustainability mandates often produce implementation gaps where narrative outpaces actual financial additionality.

The EU’s Global Gateway strategy illustrates this dynamic. Launched in 2021 as a geopolitical infrastructure investment strategy, it repositioned existing EU development cooperation pipelines under a ‘green transition’ narrative. However, quantitative analyses reveal limited new capital allocation, suggesting a relabelling of existing commitments (Bilal and Teevan, 2024). Detecting whether this shift reflects a genuine semantic evolution in operational definitions or merely a strategic redeployment of terms requires computational methods capable of resolving fine-grained sense distinctions over time.

To assess whether the semantic evolution in EU project documents indicates genuine change in meaning or strategic ambiguity, we apply our definition generation pipeline (with both SFT-only and DPO models) to the EIB corpus. We qualitatively evaluate both the scalar detection of change and the granularity of the generated definitions from a domain-expertise perspective, comparing the mod-

els across two primary axes: the base model (Llama vs. OLMo-3) and the training method (SFT vs. DPO). The models’ output was first pre-filtered using APD, as described in Subsection 5.4.

6.1 Validating Change Detection via APD

We first evaluate the alignment of the APD scores with expected policy shifts. The APD ranking consistently separates stable institutional vocabulary from highly contested policy terms. Terms expected to remain stable, such as ‘public procurement’, ‘private company’, and ‘environmental NGO’, are correctly assigned low APD scores (e.g., < 0.09). Conversely, the terms subjected to recent regulatory shocks are assigned high APD scores. For example, ‘environmental documents’ exhibits significant semantic distance when comparing 2016 and 2021 occurrences. This computationally detected shift aligns with the 2020 publication of the EU Taxonomy, which fundamentally altered the regulatory requirements for environmental documentation. Similarly, emerging technologies like ‘new hydrogen’ return high APD scores (0.391), accurately reflecting ongoing debates regarding sustainable fuel classifications. Based on domain validation, we observe an empirical APD threshold of approximately 0.27 separating substantive policy shifts from baseline syntactic variation. Only terms with the APD change score exceeding this threshold were qualitatively analyzed in the subsections below.

6.2 Sense Granularity and Preference Learning

While scalar metrics detect change, definition generation explains its nature. To evaluate this, a domain expert analysed a subset of 170 candidate phrases with sufficient temporal coverage (1,666 candidate pairs), coding each generated definition against a baseline sense as unchanged, hallucinated, or into one of three inductively derived categories of semantic shift (detailed in Section 6.3): finan-

cialisation, normative stretching, or regulatory absorption. 53% of pairs show no meaningful change, and 15% could not be reliably compared due to mismatched phrase boundaries. Among the remaining pairs, 21% are hallucinations, 20% reflect the reappearance of a previously attested sense (flagged as reintroductions rather than novel gains), 19% financialisation, 16% normative stretching, 12% regulatory absorption, and 12% combine multiple types. Qualitative inspection reveals stark differences in how models manage this sense granularity. Evaluating the training method axis, the LlamaDictionary-en_ML38BI SFT baseline suffers from excessive sense granularity. It frequently generates entirely new definitions (senses) for minor syntactic variations rather than genuine semantic shifts, producing 130 distinct senses for 'quality management' and differentiating 'protecting the environment' from 'synonym of environmental organisation' for 'environmental NGO'. Furthermore, it generates hallucinations for standard technical terms, such as linking 'necessary activity' to the US Constitution. Applying DPO to the LlamaDictionary-en_ML38BI model reduces this granularity and, when paired with clustering, consolidates terms into more stable senses while mitigating some hallucinations. Evaluating the architecture axis, the OLMo-3 models produce much lower sense granularity than Llama. OLMo-3 SFT performs exceptionally well in structural coherence and exhibits the lowest hallucination rate of all tested models. OLMo-3 DPO achieves the lowest granularity, successfully reducing redundancy. However, as detailed below, this structural advantage comes at a severe cost to domain utility.

6.3 Tracking Semantic Shifts in Policy Implementation

Crucially, the LlamaDictionary-en_ML38BI DPO model exhibited the highest sensitivity to policy-relevant nuance, successfully mapping strategic semantic expansion in the EIB corpus where other models failed. As no established typology fits this institutional-finance context, the domain expert inductively derived three trajectories of semantic shift from the generated definitions, informed by the literature on strategic ambiguity: the financialisation of physical metrics, the normative stretching of terms to absorb ecological premiums, and the regulatory absorption of external mandates into existing vocabulary. These shifts manifest as follows:

- **Financialisation:** 'additional CO2 emissions' shifted from a strict physical sense in 2014 ('additional carbon dioxide') to a financial one by 2016 ('additional cost of carbon dioxide'); similarly, 'average lifetime' moved from describing physical objects in 2019 to financial entities by 2022 ('the life of an asset').
- **Normative stretching:** 'broad participation' expanded from a purely quantitative baseline in 2017 ('participation by a large number of people') to an explicitly political one by 2022 ('involvement of as wide a section of a community as possible'). Furthermore, 'available techniques' showed a generic manufacturing practices sense in 2024 but one about methods specifically mitigating environmental damage in 2017.
- **Regulatory absorption:** 'ambitious target' shifted from a generic sense in 2015 ('a challenging goal that is striven for') to an EU-specific one by 2020 ('a goal to cut a particular environmental pollutant by a specific date').

Phrases can also acquire more than one shift type at once: 'ambitious target' independently developed a normative reading by 2023 ('an ambitious goal to reduce pollution or increase in environmental quality') alongside its regulatory one, without losing its original generic sense, which illustrates the strategic ambiguity that institutions rely on to satisfy conflicting mandates.

6.4 Observed Limitations in Definition Generation

Qualitative analysis exposes a clear trade-off between structural granularity and domain sensitivity across both comparison axes. First, comparing architectures, the OLMo-3 models sacrifice critical policy sensitivity to achieve their low granularity. While Llama models capture nuanced shifts in terms like 'additional CO2 emissions', OLMo-3 models generate overly generic, dictionary-style definitions that fail to capture institutional context. Second, comparing training methods within architectures reveals diverging effects. For Llama, DPO is a necessary intervention to make the highly granular SFT outputs usable, preserving policy sensitivity while smoothing structural noise. Conversely, applying DPO to OLMo-3 damages its qualitative performance. While OLMo-3 SFT maintains some

alignment with the EIB context and minimizes hallucinations, OLMo-3 DPO collapses specialised terms entirely. For instance, OLMo-3 DPO hallucinates 'new hydrogen' as a 'Japanese fuel' and produces definitions entirely detached from the political economy context. This indicates that while preference optimisation improves highly granular models like Llama, it excessively penalises the domain-specific jargon required for institutional analysis in models like OLMo-3.

6.5 Bridging Detection and Interpretation

Our qualitative assessment shows that preference-optimised definition generation successfully translates abstract semantic change scores into explicit, human-readable evidence. While scalar metrics flag that a change has occurred, our pipeline maps these contextualised occurrences to coherent definitions-as-senses, providing political scientists and sustainability researchers with a scalable tool to track how operational terms related to green investments evolve across institutional disclosures. Ultimately, by complementing change scores with these structured definitions, researchers and policymakers can systematically test claims such as the Global Gateway's 'green transition' narrative against project-level operational vocabulary, distinguishing genuine change in meaning from an institution's strategic expansion of core concepts, checkable at corpus scale rather than through manual document review.

7 Conclusion

We have shown that preference optimization (in the form of DPO) may help to overcome the problem of too granular senses in the outputs of existing definition generation models. This in turn increases the performance of semantic change detection systems in the 'definitions as senses' paradigm.

We experimented with different strategies of merging definitions-as-senses where they were still too granular and found that it is largely not necessary for modern definition generators. Interestingly, a large crowd-sourced dictionary (DBnary) allows to train more powerful models than datasets derived from smaller dictionaries created by professionals. Another practically useful finding is that a 7 billion parameter model performs on par with a 8 billion parameter one, although it is not possible to name a reason for it from our study (the models are also different in their pretraining and instruction-tuning

data, tokenizers etc.).

While the gain in reducing number of senses may look minor in absolute numbers compared to the computational cost of DPO, for a human end user it means much less (may be hundreds) definitions to read and analyse, so we still believe DPO is useful. Preference optimization has to be done only once and we have already done it. Any social scientist will be able to use the models we publish.

We applied our pipeline to a real-life task of studying diachronic change of sustainability concepts in EU project documents. Applied to the EIB corpus, the pipeline made visible how the Bank's own operational vocabulary is reclassified over time to adapt to external political and regulatory mandates. It offers concrete evidence that a definition generator retaining domain-specific nuance, as achieved here by LlamaDictionary-en_ML38BI DPO, can make an institution's narrative shifts traceable directly in its operational documents, not only its curated public statements.

Limitations

A potential risk of our work is that we did not check DBnary for offensive content (while Wiktionary is moderated, it can be edited by anyone); we also did not check the base models for the ability to generate such a content.

Our experiments with clustering are limited to only one sentence embedding model and to only one algorithm. However, we believe that the future work should concentrate on developing better definition generators rather than on dataset-specific post-processing of the definitions.

We do not use any thresholding for APD. Thus, our current pipeline does not map the exact values of APD to the 'real' change/not change labels. This decision is currently left to a human expert.

Impressions of a human expert do not exactly align with the models' performance on a curated benchmark (featuring only 37 words, all of them belonging to the basic vocabulary of the English language). The problem of LSCD models performing differently on curated benchmarks and 'in the wild' has also been reported in the prior work (Umarova et al., 2025). Thus, there is a need for larger and domain-specific benchmarks. Qualitative categorisation in Section 6 was conducted by a single domain expert; adding independent annotators and reporting inter-rater agreement is a natural next step for this line of work.

Acknowledgments

The computations were performed on resources provided by Sigma2 - the National Infrastructure for High-Performance Computing and Data Storage in Norway.

This project has received funding from the European Union's Horizon Europe Marie Skłodowska-Curie Actions Doctoral Network under grant agreement No 101120060.

References

- Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.
- San Bilal and Chloe Teevan. 2024. [Global Gateway: Where now and where to next?](#) Technical Report Discussion Paper No. 368, ECDPM.
- Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. [An open dataset and model for language identification](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.
- Pierluigi Cassotti, Stefano De Pascale, and Nina Tahmasebi. 2024. [Using synchronic definitions and semantic relations to classify semantic change types](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4539–4553, Bangkok, Thailand. Association for Computational Linguistics.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). Preprint, arXiv:2402.03216.
- Steve Connelly. 2007. [Mapping Sustainable Development as a Contested Concept](#). *Local Environment*, 12(3):259–278.
- European Investment Bank. 2015. [Guide to accessing environmental and social information/documents held by the EIB](#). Technical report, European Investment Bank.
- Mariia Fedorova, Nikolay Arefyev, Maja Buljan, Jindřich Helcl, Stephan Oepen, Egil Rønningstad, and Yves Scherrer. 2026. [OpenLID-v3: Improving the precision of closely related language identification – an experience report](#). In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 275–292, Rabat, Morocco. Association for Computational Linguistics.
- Mariia Fedorova, Andrey Kutuzov, Francesco Periti, and Yves Scherrer. 2025. [Explaining novel senses using definition generation with open language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22294–22302, Suzhou, China. Association for Computational Linguistics.
- Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. 2024. [Definition generation for lexical semantic change detection](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5712–5724, Bangkok, Thailand. Association for Computational Linguistics.
- Peer C. Fiss and Edward J. Zajac. 2006. [The Symbolic Management of Strategic Change: Sensegiving via Framing and Decoupling](#). *The Academy of Management Journal*, 49(6):1173–1193.
- 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.
- 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, 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.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- 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.
- Leonard Kaufman and Peter J Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.
- Andrey Kutuzov, Mariia Fedorova, Dominik Schlechtweg, and Nikolay Arefyev. 2024. [Enriching word usage graphs with cluster definitions](#).

- In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6189–6198, Torino, Italia. ELRA and ICCL.
- Shirley Leitch and Sally Davenport. 2007. [Strategic ambiguity as a discourse practice: the role of keywords in the discourse on ‘sustainable’ biotechnology](#). *Discourse Studies*, 9(1):43–61.
- Vladimir I Levenshtein and 1 others. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.
- Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. 2020. [Analyzing Sustainability Reports Using Natural Language Processing](#). *arXiv preprint*.
- 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.
- Stephan Oepen, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie V. Burchell, Lucas Georges Gabriel Charpentier, Pinzhen Chen, Mariia Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindrich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Bhavitvya Malik, Vladislav Mikhailov, Amanda Myntti, and 10 others. 2026. [HPLT 3.0: Very large-scale multilingual resources for LLMs and MT. Mono- and bi-lingual data, multilingual evaluation, and pre-trained models](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 1409–1434, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2025. [Olmo 3](#). *Preprint*, arXiv:2512.13961.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Francesco Periti, David Alfter, and Nina Tahmasebi. 2024. [Automatically generated definitions and their utility for modeling word meaning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14008–14026, Miami, Florida, USA. Association for Computational Linguistics.
- Francesco Periti, Roksana Goworek, Haim Dubossarsky, and Nina Tahmasebi. 2025. [Definition generation for word meaning modeling: Monolingual, multilingual, and cross-lingual perspectives](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26004–26024, Suzhou, China. Association for Computational Linguistics.
- Bach Phan-Tat, Kris Heylen, Dirk Geeraerts, Stefano De Pascale, and Dirk Speelman. 2026. [Evaluating the evaluator: Problems with semeval-2020 task 1 for lexical semantic change detection](#). *Journal of Open Humanities Data*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2024. [Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication](#). *Finance Research Letters*, 61:104979.
- 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.
- Dominik Schlechtweg, Sachin Yadav, and Nikolay Arefyev. 2025. [The LSCD benchmark: a testbed for diachronic word meaning tasks](#). *Preprint*, arXiv:2404.00176.
- Gilles Sérasset. 2012. [Dbnary: Wiktionary as a LMF based multilingual RDF network](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2466–2472, Istanbul, Turkey. European Language Resources Association (ELRA).
- Maksim Tatarinov and Aleksandr Demidovsky. 2025. [Interpretable approach to detecting semantic changes based on generated definitions](#). In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2025”*.
- Khonzoda Umarova, Lillian Lee, and Laerdon Kim. 2025. [Current semantic-change quantification methods struggle with discovery in the wild](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35354–35367, Suzhou, China. Association for Computational Linguistics.
- Sachin Yadav and Dominik Schlechtweg. 2025. [XL-DUREl: Finetuning sentence transformers for ordinal word-in-context classification](#). In *Proceedings*

of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 338–351, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

A APD descriptive statistics

Table 5 presents descriptive statistics of APD calculated by XL-Durel for the ESDS data.

	APD value
mean	0.13
std	0.10
min	0.0
25%	0.07
50%	0.11
75%	0.17
max	0.72

Table 5: Descriptive statistics of APD calculated by XL-Durel.

B Examples of fine-tuning instances

B.1 SFT

An example of a SFT training instance (used for fine-tuning the OLMo base model; we did not perform own SFT of LlamaDictionary-en_ML38BI, but reused the prior work).

- Target word: Suffer
- Prompt + usage example: If you keep partying like this, your school-work will suffer. What is the definition of suffer?
- Gold prediction: To become worse.

B.2 DPO

The next example shows a DPO training instance, created from LlamaDictionary-en_ML38BI SFT predictions:

- Target word: Tip
- Prompt + usage example: Please provide a concise definition for the meaning of the word "Tip" in the following sentence: (Dirks was an investment analyst who learned from a former employee of Equity Funding that the company had been fraudulently manufacturing insurance policies. Dirks tipped several institutions

which then liquidated \$16 million in Equity Funding stock before the fraud was exposed and the bottom fell out of the market. According to the Court, Dirks was a hero for (eventually) exposing the fraud. The SEC’s censure of Dirks for tipping inside information was reversed because, the Supreme Court held, the liability of a tippee derives from that of the tipper. If the tipper is without sin, so is the tippee. Here, the employee’s purpose in informing Dirks was to expose the fraud, so the tipper breached no duty. Moreover, according to the Court, for there to be a breach of duty, there must be personal benefit (in the form of pecuniary gain or reputational benefit that will translate into future earnings) arising to the tipper from the disclosure

- Chosen: To give a piece of private information to; to inform (someone) of a clue, secret knowledge, etc.
- Rejected: To give a piece of private information to; to inform (someone) of something not widely known.

In this instance, the chosen definition was generated from the shown usage example ‘(Dirks was...’, and the rejected definition was generated from the usage example ‘If the institutional trading patterns we observe are a result of tipping, rather than precipitated by some other event, then we should see only a slight increase in the number of institutions active in the market (tipping would precipitate entry by, at most, the few institutions that were tipped). At the same time, given the nature of the reports, we should see an increase in the average buying activity of institutions (we cannot identify the specific institutions that were tipped, so we can only look at averages) and see little change in selling’. The original gold definition for both usages was ‘To give a piece of private information to; to inform (someone) of a clue, secret knowledge, etc. [from late 19th c.’.