A Sense-Level Benchmark for Denotational and Connotational Meaning Relations

Pierluigi Cassotti, **Naomi Baes**, **Jader Martins Camboim de Sá**, **Francesco Periti**, **Stefano De Pascale**, **Nick Haslam**, **Dirk Geeraerts**, **Nina Tahmasebi**

University of Gothenberg, The University of Melbourne, KU Leuven, Vrije Universiteit Brussel

Luxembourg Institute of Science and Technology

{pierluigi.cassotti,nina.tahmasebi}@gu.se, {nbaes,nick.haslam}@unimelb.edu.au

jader.martins@list.lu,{francesco.periti,stefano.depascale,dirk.geeraerts}@kuleuven.be

Abstract

Polysemy enables a single word to convey multiple related meanings, reflecting the conceptual and emotional aspects of language evolution. We introduce the first sense-level benchmark for modeling semantic relations between word senses, uniting denotational and connotational aspects of meanings. The benchmark distinguishes denotational relations, such as generalization or metaphor, as well as two connotational dimensions: valence and arousal. We evaluate large language models (LLMs), GPT-4o, Llama 3.1, and DeepSeek, in zero-shot and fine-tuned settings. Results show that GPT-4o best aligns with human affective judgments, while a fine-tuned RoBERTa model excels at classifying denotational relations.

1 Introduction

Polysemy, that is, the use of a single word to express multiple related meanings, can be seen as a strategy of linguistic economy (Zipf, 1949). By reusing existing lexical items, speakers reduce both cognitive and communicative effort while keeping language expressive and adaptable. The extension of a word's meaning to new contexts and concepts does not occur randomly. Instead, a word's senses are **systematically connected through cognitive mechanisms** such as similarity and contiguity (Blank, 1997; Schlechtweg, 2023). By projecting familiar concepts into new contexts, speakers construct intricate semantic networks that sustain polysemy and guide meaning change (Brochhagen et al., 2023). These semantic relations not only organize the lexicon but also reveal the cognitive and cultural forces driving language evolution.

The emergence of new word meanings, a process called semantic change, is a gradual one. Geeraerts (2020) describes this development in two stages: first, a word's sense is extended to novel contexts (*context variance*), and second, the new usage becomes established as a distinct meaning (*semantic change*). We go beyond the state of the art and investigate how new senses emerge and connect by simultaneously distinguishing two fundamental levels of meaning: denotational (i.e., conceptual meaning) and connotational (i.e., the emotional value). We thus simultaneously consider *types* of denotational change and *dimensions* of connotational change. Relying on the observation that most synchronic sense relations have arisen from diachronic change, we propose a framework that integrates connotational and denotational meaning shifts to capture both the cognitive underpinnings and the diachronic trajectories of semantic change.

Original Contributions: The present study is the first to (1) investigate, computationally, the interaction of connotational and denotational aspects of meaning that are present at the sense-level; (2) release human-annotated benchmarks for modeling semantic relations between word sense pairs annotated for change types, dimensions, and their degree of change; (3) evaluate the ability of LLMs to model the relations between word senses by means of pairs of sense definitions, pairs of usages, and both. Finally, to showcase the ability to scale the studies, we (4) tests whether denotational types map onto connotational dimensions (and if so, which) and conduct a dictionary-wide study exploring correlations between those dimensions and types. Our long-term aim is to apply these systems for diachronic analysis, to study the emergence of new meanings as they arise. All code and prompts will be released on Github, datasets on Zenodo, and models on Huggingface upon acceptance.

2 Types and Dimensions

In this paper we investigate the semantic relations between sense pairs that belong to the same word. The theoretical assumption is that synchronic relations result from diachronic processes. Therefore, by identifying synchronic relations, we aim to cap-

Word	Original sense	New Sense(s)	Denotational label	Connotational label
trauma	Physical injury (physical wound)	An emotional wound or shock that may have long-lasting effects	Metaphorical extension	*arousal*
plane	Flat surface (geometric sense)	Level of existence, consciousness, or development (e.g., "higher plane")	Metaphorical extension	–
geek	Socially awkward or odd individuals	People with deep expertise or passionate interest in a specific field	Generalization	*valence*

Table 1: Examples of words with their original and new senses as well as denotational and connotational changes.

ture the corresponding diachronic processes in corpus data. Therefore, *relation types/dimensions* and *change types/dimensions* are used interchangeably unless otherwise stated.

Largely due to their development within separate research traditions, the denotational and connotational levels of meaning are often operationalized in different ways. Denotational meaning, typically studied in linguistic research, is often described in separate and mutually exclusive types, of which we consider five, following Ullmann (1957); Geeraerts (1997). Connotational meaning, by contrast, was conceptually formalized in linguistics and philosophy but has been empirically investigated most extensively in psychology, where it is typically measured as dimensions using continuous affective scales (Osgood et al., 1975; Russell, 2003).

Following recent work on detecting types of semantic change (Cassotti et al., 2024), we annotate on the level of pairs of definitions. However, to further provide contextual cues, we also follow Schlechtweg et al. (2020) in annotating usage pairs (i.e., pairs of sentences). While we will adhere to categorical labels for denotational types, we use ordinal labels for connotational dimensions. Each word sense is rated with a *valence* and *arousal* value of *low, neutral, high*, and differences are derived between sense pairs using a three-point scale: `0 = unchanged`, `1 = small difference`, `2 = big difference`. See Table 2 for more details.

Table 1 illustrates words that have experienced denotational changes, two of which have an associated connotational change. The word *trauma*, for example, has two senses, where one is a metaphorical extension of the other (from a physical wound to an emotional one), whereby its second sense can be considered lower on the *arousal* dimension.

2.1 Denotational Types

We focus on the following types of denotational relations, drawing on previous work (Blank, 1997; Cassotti et al., 2024).

`generalization`: the old meaning is a sub-case of the new meaning (also known as *broadening* or *widening*). For example, *calf*, which originally meant 'the young of the domestic cow' now also means 'the young of various large mammals (e.g. elephants, whales)'.

`specialization`: the new meaning is a sub-case of the old meaning (also known as *narrowing* or *restriction*). An example is *escort*, which used to refer broadly to 'persons accompanying another to give protection, provide supervision, or as a courtesy' but has since acquired a more restrictive interpretation as 'a sex worker hired to go with someone to a social event'.

`homonymy`: the new meaning is etymologically unrelated to the old meaning (and the meanings belong to different lexemes); For example, *bull* is both a continuation of the Old Norse *bole* 'a male bovine' and of the Latin *bulla* 'a solemn, sealed papal letter'.

`metonymy`: the referent of the new meaning is related in terms of contiguity to the referent of the old meaning; An example is *head*, which first-most refers to 'the upper division of the animal body that contains the brain' and afterwards extended its meaning to refer to 'the seat of the intellect: mind'.

`metaphor`: the referent of the new meaning is related in terms of figurative similarity to the referent of the old meaning. An example is *wing*, which first-most refers to 'one of the movable feathered paired appendages by means of which a bird is able to fly' and afterwards extended to reference 'a part or feature of a building usually projecting from and subordinate to the main or central part'.

Because specialization and generalization depend on the point of view (i.e., which sense is used as the starting point), we will also consider them jointly and refer to them as *taxonomical* relations.

2.2 Connotational Dimensions

In addition to the types of denotational relations, this study focuses on two universal dimensions of affective meaning, *valence* (good–bad) and *arousal* (calm–excited) (Russell, 2003), which have been linked to connotational meaning (Baes et al., 2024).

`valence`: the degree of pleasantness or unpleasantness associated with a word's meaning, reflecting its position along a positive-negative evaluative continuum. It is captured in affective lexica (Jurafsky and Martin, 2025) such as ANEW (Bradley and Lang, 1999), the Warriner norms (Warriner et al., 2013), and NRC-VAD (Mohammad, 2018), where words like *murder* are rated as highly negative, *object* as neutral, and *love* as highly positive.

`arousal`: the degree of energy, stimulation, or alertness conveyed by a word's meaning, corresponding to an excited-calm (or active-passive) continuum. It is also captured in affective norms (Bradley and Lang, 1999; Warriner et al., 2013; Mohammad, 2018) (excited–calm or active–passive dimension), where words like *dull* are rated as low-*arousal*, *object* as neutral, and *insanity* as high-*arousal*.

3 Human annotation

One of our primary contributions is the creation of a manually annotated dataset comprising word senses and sense pairs, which we later employ to evaluate computational models. While we use existing datasets for *training* models on types of semantic change, all models are also *evaluated* on data newly produced through expert annotation.

3.1 Denotational types

Currently, there are no databases or dictionaries for English that report the types of denotational meaning relations surveyed in this study. We therefore take as a starting point the Dutch online dictionary *Algemeen Nederlands Woordenboek* (ANW; Dutch Language Institute n.d.) which provides annotations of such semantic relations. There are two reasons to consider this a relevant resource. First, the two linguists authoring this paper are Dutch native speakers, which ensures reliability in identifying semantic nuances. Second, given that Dutch and English are closely related Germanic languages, we can assume that the semantic relations found in the Dutch lemmas are likely to be found in their English translation, obtained through Open Multilingual WordNet (Bond et al., 2016). The data creation involves using the ANW as a 'seed dictionary'. For instance, if the ANW reports a relation of *generalization* between the readings 'appealing to a higher court for review of a judgment or decision' and 'urgent request' for the Dutch lemma *appel*, we looked up whether the corresponding English lemma *appeal* would show a similar relation in its polysemy, based on the entry of that lemma in the online Merriam-Webster Dictionary (MWD). By extracting the respective definition pairs from the English lemma in MWD we were able to create a final evaluation set for English that contains 60 pairs of senses for each of the five denotational relations above. In total, 300 sense pairs were annotated. The compilers of the dataset, who are also the linguistic authors trained in lexical semantics, followed a two-round procedure whereby the 60 pairs of a type are harvested by one compiler, checked by the other and, if needed, complemented with new pairs to arrive at a shared revised set.

3.2 Connotational dimensions

For annotating connotational dimensions, no pre-existing seed dictionaries are available. Therefore, we create one and use computational modeling as a filtering step before reaching our final dataset. Starting from a balanced sample of 1,905 human-annotated seeds[1] we trained RoBERTa[2] to classify WordNet glosses on their *arousal* level (low, neutral, high). This yielded an extended dataset of 23,997 senses automatically annotated for *arousal*. From this set, we then selected 250 sense pairs with the largest model-predicted differences in arousal and 250 with the largest differences in valence for expert annotation, conducted blind to the model's classifications, resulting in a final dataset of 652 unique synsets and 952 annotated sense pairs.

Because annotating pairwise differences in connotational dimensions between senses proved highly challenging for humans (see App. B) and LLMs (see Table 4), we instead annotate each individual sense from the 952 sense pairs for both *valence* and *arousal*, and then computed their differences as the absolute value between the two ratings. Drawing primarily on sense definitions

[1]These seeds contained senses from only a few words, and the annotation revealed few instances where there was a difference in either dimension, thus we opted to use the seeds for training the classifier to create an extended dataset.

[2]FacebookAI/RoBERTa-base `https://huggingface.co/FacebookAI/RoBERTa-base`

Condition	Setting	Task (Dimension Specific)	Scale (*valence*, *arousal*)
Instance	Single definition	Rate one sense definition on levels	Low, Neutral, High
	Single sentence	Rate one sentence example of a sense on levels	Low, Neutral, High
	Definition & Sentence*	Rate one sense sentence usage for levels	Low, Neutral, High
Difference	Pair of definitions	Compare two sense definitions	Same, Small, Big
	Pair of sentences	Compare two sentence examples	Same, Small, Big
	Pair of definitions & sentences	Compare pairs of sense definitions & sentences	Same, Small, Big

Table 2: Experimental settings for evaluating sense relations between dimensions. *Note:* * = Human annotation.

and supplementing with sentence examples, each sense was independently annotated by two psychology scholars for *valence* and *arousal* values on a discrete ternary scale (low = -1, neutral = 0, high = +1). Reliability was moderate for *arousal* ($\rho = .60$, $p < .0001$; $\kappa = .52$–$.58$ weighted; $n = 662$), with most disagreements involving the "low/neutral" classes. For *valence*, the reliablity was moderate to strong ($\rho = .83$, $p < .0001$; $\kappa = .74$–$.83$; $n = 662$), with disagreements mostly concerning the "high-/neutral" classes. The final benchmark datasets contain 952 sense pairs (662 senses, 235 unique words), each annotated for levels of *valence* and *arousal*. Human and LLM instructions are similar.

4 Experimental Setup

This section introduces the experimental setup for evaluating LLMs on two tasks: determining the degree of *arousal*/*valence* inherent in senses and between their pairs, and classifying the denotational types of semantic relation between senses. In both cases, prompts were engineered with domain experts and optimized as required. We consider both open models (such as Llama 3.1 8B, Llama 3.1 70B, and DeepSeek) and closed models like GPT-4o. We evaluate the LLMs using zero-shot prompting (see App. A for prompt design). Additionally, for the relation types, since large-scale datasets are available from various sources (see App. D), we gather them and fine-tune a Llama 8B-model and train a RoBERTa-large–based classifier. More information on generation parameters, model training, and hyperparameters can be found in App. E.

4.1 Denotational types

Given a pair of definitions of a word, we want to determine the relation between the word senses. We use the five types of relations defined in Sec. 2. Since the fine-tuning dataset (WN+CN+UM) described below also allows for the study of *auto-antonymy*, we will include this class among the types that we model computationally (but for which we do not have any human annotated data because the class was not present in the seed dictionary, and thus *auto-antonyms* are excluded from Figure 1).

`(auto-)antonymy:` also called *contronymy* or *enantiosemy*, is when a word develops a new meaning that expresses a contrast, or is in opposition, to its old meaning. For instance, *dust* both means 'to make free of dust' and its opposite 'to sprinkle in the form of dust'.

Zero-shot For zero-shot approaches, the prompt was designed by a linguist, and the model is provided with a pair of definitions and instructed to choose one of the relations mentioned above.

Finetuning For the relation types, we unified existing datasets, and used them to train a classifier based on RoBERTa-large as well as to fine-tune the Llama 3.1 8B model. For *metaphor, metonymy,* and *homonymy*, we used the ChainNet dataset (Maudslay et al., 2024), which extends WordNet by labeling these relations between pairs of synsets (and thus between pairs of definitions). We further extended ChainNet using UniMet, a dataset of metonymies (Khishigsuren et al., 2022). For *generalization* and *specialization* we used the datasets introduced by Cassotti et al. (2024), which also makes use of WordNet synsets. In this case, the definition pairs belong to connected hypernym and hyponym synsets (for example, a generalization pair is given by coupling the definition of the hyponym *dog* with the definition of the hypernym *animal*). The rationale behind this choice is that extensive training datasets for these relations types do not exist; therefore, relations between synsets, taken from a lexical database, are used as a proxy for relations between senses. The final dataset obtained by concatenating these datasets (i.e., WN+CN+UM) was then split into training, development, and test sets. See Table 6 for dataset statistics.

4.2 Connotational dimensions

To evaluate the ability of LLMs to annotate the *valence* and *arousal* of senses, we (1) drew on psychology expertise to design task-specific instruc-

tions for humans and LLMs, ensuring close resemblance between prompts and guidelines; (2) validated LLM performance on the human-labeled subset via inter-annotator agreement; (3) conducted prompt optimization to maximize alignment between LLM-human judgements (see App. A); and (4) used the best-performing prompt for evaluation.

For humans, the task involves annotating senses for their levels of the dimensions reflecting connotational meaning: *valence* and *arousal* (low, neutral, high), using sense definitions and sentence examples. For LLMs, the tasks are outlined in the six experimental settings shown in Table 2. The instance conditions (Inst) comprise of (I) single definition, (II) single sentence, or (III) single definition and sentence settings, where LLMs annotate either a dictionary definition, a sentence, or both, containing the target word. Out of the 952 annotated pairs, we compare the models only on the pairs where both annotators agree (for valence: 508; arousal: 369). In the pairwise difference conditions (Diff), LLMs assign same, small difference, big difference to each pair of sentences. There are three Diff conditions: the comparison of (IV) two definitions or (V) two sentences for the difference in *valence* and *arousal*. In the final (VI) pairwise difference sentences and definitions setting, they compared pairs of definitions and usages and judge the magnitude of the difference in *valence* and *arousal*.

4.3 Affective Sign-Shift Analysis

To showcase the possibilities of classifying and cross-relating denotational types and connotational dimensions, we examine whether different types of semantic relations (*taxonomical*, *metonym*, *metaphor*, *antonymy*, and *homonymy*) show distinct patterns of connotational change between the two senses of a word. Because *valence* and *arousal* values were encoded on a discrete ternary scale $\{-1, 0, +1\}$, affective divergence was operationalized as a *sign shift*, that is, whether the two senses had received a different value on that scale (e.g., a positive $\rightarrow$ negative shift in *valence*, or a positive $\rightarrow$ neutral shift in *arousal*). For each synset pair S_1, S_2, *valence* and *arousal* values $S_1 = (v_1, a_1)$ and $S_2 = (v_2, a_2)$ were retrieved and compared by computing $\Delta v = |v_2 - v_1|$ and $\Delta a = |a_2 - a_1|$. A sign flip was defined whenever Δv or $\Delta a = 2$ (big difference on the 3-point sense difference scale) indicating a transition across the neutral midpoint while $\Delta = 1$ (or small differences) means that one of the values is neutral while the other is ± 1 and $\Delta = 0$ (or unchanged) means that both have the same value.

Model	WN+CN+UM	Denot. Dataset
GPT-4o	0.522	0.680
DeepSeek-V3.2-Exp	0.426	0.487
Llama 3.1 8B Instruct	0.286	0.287
Llama 3.1 70B Instruct	0.445	0.580
RoBERTa-large	0.734	**0.684**
Llama 3.1 8B FT	**0.737**	0.658

Table 3: Weighted F1 scores on the WC+CN+UM test set and the denotational dataset.

5 Results

In this section, we present the results of the LLM evaluations for dimensions and types.

5.1 Denotational types

The results, reported in Table 3, show a clear contrast between LLMs evaluated in a zero-shot setting and models that were fine-tuned specifically for the task of classifying types of semantic relations.

On the WN+CN+UM dataset, the best-performing systems (F_1^{weighted}) are RoBERTa-large (0.734) and Llama 3.1 8B FT (0.737). These two models were trained on the task, unlike GPT-4o, DeepSeek, or Llama Instruct, which were evaluated zero-shot. On the denotational dataset in our benchmark, containing expert annotated sense pairs, the picture shifts slightly: RoBERTa-large achieves the best score (0.684), with GPT-4o close behind (0.680). This suggests that, while fine-tuning on proxy datasets helps on synthetic or mixed resources, on a linguistically curated gold-standard test set, a traditional transformer like RoBERTa-large remains very strong, even outperforming GPT-4o. Smaller LLMs without fine-tuning (Llama 8B Instruct, DeepSeek) lag behind substantially, showing that **general-purpose models still underperform on lexical-semantic classification when compared to dedicated models**.

Examining the confusion matrix (Figure 1) for RoBERTa-large on the denotational dataset reveals patterns in how the model handles the five types of denotational change. The results highlight a fundamental challenge in modeling sense relations: taxonomic relations (*generalization/specialization*) are systematically confused with each other, while figurative relations (*metaphor/metonymy*) are harder to disentangle and are often confused with taxonomic or unrelated categories. *Homonymy* remains especially challenging as it requires distinguishing true

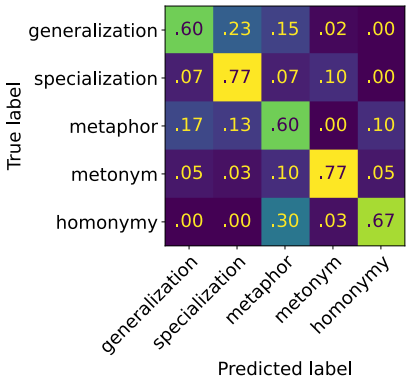

Figure 1: Confusion matrix denotational dataset - RoBERTa-large.

unrelatedness from distant figurative extensions.

In sum, the patterns of confusion captured in Figure 1 are expected along the lines of the two main drivers of semantic change: similarity and contiguity. *Generalization, metaphorical extension* and *homonymy* can be placed on a cline of decreasing similarity between the senses: from expansion within a category (for *generalization*), to jump across categorical domains (for *metaphor*) lacking any relatedness motivation (for *homonymy*). Notably, the cross-confusion scores between those three relations reflect a similar cline: *generalization* has high confusion (15%) with *metaphor*, but none with *homonymy*; *metaphor* is both confused with *generalization* (17%) and *homonymy* (10%), and, last, *homonymy* is greatly confused with *metaphor* (30%) but not with *generalization*.

Conversely, *metonymy*, is commonly understood in terms of contiguity, and stands out among the relations. However, for *specialization*, which implies a restriction of the category, both the higher score compared to metonymy and the low confusion are surprising and deserve closer study in the future.

5.2 Connotational dimensions

The results, reported in Table 4 on the pairs of senses where both annotators agree (508 for valence and 369 for arousal), show a clear pattern across models, experimental settings, and affective dimensions. For model evaluation, we excluded cases where we were not able to parse the model's answer and filtered out examples where the target word was not contained in the example sentence (a situation that can occur in WordNet when a synonym is used instead of the target word). Overall, *valence* is consistently easier to model than *arousal*, demonstrated by higher correlations between LLM and human judgments. **Using the sentence and the definition (*sent+def*)** when prompting LLMs to annotate dimensions of connotational meaning **produces the strongest correlations with human judgments**. This is unsurprising, as the same approach was used for human annotation, and the richer context helps to better evaluate affective meaning. Using only *definitions* also improves performance compared to using only *sentences*, confirming that more explicit semantic framing (that more closely resembles the prototypical meaning of a word) makes it easier for LLMs to align with affective human evaluations.

When comparing the two strategies, Inst (where the model gives individual *valence*/*arousal* scores and the difference is computed afterwards) tends to outperform Diff (where the model is asked to directly judge the magnitude of the difference), particularly for *valence*. The advantage of Inst is consistent with the idea that labeling each sense individually provides a more stable basis for comparison. However, for *arousal*, the pattern is less consistent: while GPT-4o and DeepSeek generally perform better with Inst, Llama 70B achieves its strongest result with Diff.

Looking at the models individually, GPT-4o is the strongest across almost all conditions, reaching near-human performance in the sent+def Inst setting ($\rho = 0.927$ for *valence*, $\rho = 0.854$ for *arousal*). DeepSeek-V3.2-Exp is a solid second-best, consistently outperforming Llama models. The Llama 3.1 models perform worse overall, with the 70B variant generally stronger than the 8B, particularly for *valence*. Interestingly, Llama 70B surpasses GPT-4o in *arousal*-Diff, even if alignment with human affective judgments remains weaker overall.

5.3 Large-scale analysis of the interaction between connotation and denotation

We performed a vocabulary-wide analysis of sense relations by extracting sense pairs from WordNet. First, we select, among all possible lemmas in WordNet, only those for which all associated synsets have at least one example sentence containing the lemma itself, resulting in a total of 1,892 lemmas. For each lemma, we then generate all possible directional pairs by combining its synsets, meaning that for every pair $\langle$A, B$\rangle$, the reversed pair $\langle$B, A$\rangle$ is also included. Each pair is then classified for relation type using the fine-tuned RoBERTa-large model. We retain only

		Intensity		Polarity	
		Inst	Diff	Inst	Diff
sentence					
	GPT-4o	0.680	0.316	0.804	0.441
	DeepSeek-V3.2-Exp	0.608	0.294	0.783	0.295
	Llama 3.1 8B Instruct	0.277	0.153	0.565	0.181
	Llama 3.1 70B Instruct	0.354	0.281	0.671	0.275
definition					
	GPT-4o	0.824	0.335	0.902	0.633
	DeepSeek-V3.2-Exp	0.779	0.374	0.893	0.468
	Llama 3.1 8B Instruct	0.459	0.230	0.696	0.333
	Llama 3.1 70B Instruct	0.356	0.362	0.802	0.376
sent+def					
	GPT-4o	**0.865**	0.450	**0.932**	**0.636**
	DeepSeek-V3.2-Exp	0.818	0.395	0.909	0.419
	Llama 3.1 8B Instruct	0.477	0.247	0.434	0.362
	Llama 3.1 70B Instruct	0.616	**0.482**	0.842	0.429

Table 4: Zero-shot results for *arousal* and *valence* across six experimental settings, grouped by dataset (rows) and candidate models. Values show correlation (ρ) between LLM judgments and human judgments (100% agreement between 2 expert annotators).

those pairs for which RoBERTa-large predicts the same relation type in both directions. In cases of mixed predictions, such as homonymy–metaphor or homonymy–metonymy, we label the pair as metaphor and metonym, respectively. Pairs involving generalization and specialization (or both) are labeled as taxonomical. Arousal and valence scores for each sense are obtained using GPT-4o annotations, and the final dataset includes only pairs for which both arousal and valence values could be successfully parsed. Figure 2 shows the differences in classified dimensions by relation type. On the y-axis we find Δa while on the x-axis we find Δv values. The left bottom-most plot shows the number of sense pairs for each relation that do not exhibit any difference in neither *arousal* nor *valence* values. The plot above shows those for which there is a small change in *arousal* but no change in *valence*.

As expected overall, there are few pairs that have a large difference in arousal or valence, and even fewer that have a large difference in both. For both *arousal* and *valence* only, the changes occur primarily in the antonymy relation (unsurprisingly), followed by the taxonomical relations, metaphor, metonym and finally homonyms.

A slightly different pattern appears when examining the middle plot, where one of the senses in the pair is neutral (0) while the other has either positive or negative *valence*/*arousal*. In this case, there are no metonyms compared to metaphors. In general, there are more sense pairs with sign flips in *arousal* than in *valence*. The proportion of sign flips for each type is outlined in App F.

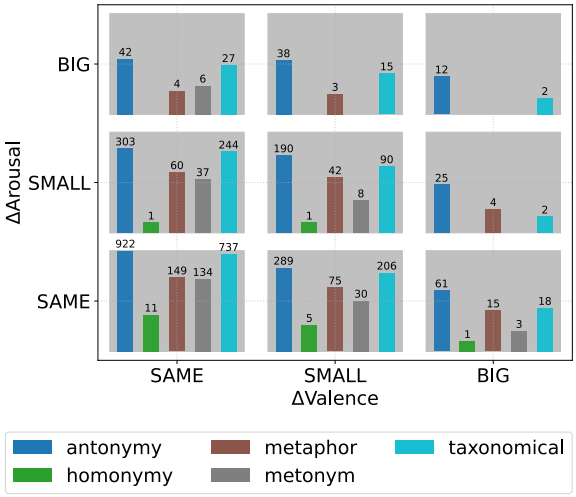

Figure 2: Distribution of sign flips, *arousal* on the y-axis and *valence* on x-axis. SAME = no change, SMALL = small change from or to neutral, BIG = large change (sign flip), from one extreme to the other.

6 Related Work

6.1 Connotational dimensions

While there are several proposed dimensions of semantic change, including Sentiment/Polarity, Intensity, Breadth/Dimension and Relation (Baes et al., 2024; de Sá et al., 2024b), the present study focuses on two universal dimensions of affect, *valence* (good–bad) and *arousal* (calm–excited) (Russell, 2003), that have been linked to two primary dimensions of connotational meaning, Evaluation ("good/bad") and Potency ("strong/weak") within the SIBling framework (Baes et al., 2024). SIBling further connects these two dimensions to types of lexical semantic change, such that each pair of change types (amelioration-pejoration and hyperbole-meiosis) corresponds to a rise or fall along a single pole (*valence*/Sentiment and *arousal*/Intensity, respectively). Notably, *valence* and *arousal* are chosen because they are psychologically universal, reliably measured across cultures, and provide a validated foundation for modeling connotational meaning in semantic change.

Nevertheless, there are no evaluation datasets with sense-labeled data for *valence* and *arousal*. Most available resources exist at the word level (Jurafsky and Martin, 2025; Bradley and Lang, 1999; Warriner et al., 2013; Mohammad, 2018) or the text level (Buechel and Hahn, 2022). Recent approaches use LLMs to generate benchmark datasets.

Yet, Baes et al. (2025)'s yields diachronic, domain-specific, sentences containing a target term (without sense information) while de Sá et al. (2024a)'s contains judgments for the Polarity of sense pairs, but yielded subpar evaluation results.

6.2 Denotational types

Historical linguistics has a long tradition of classifying mechanisms of semantic change (Blank, 1997; Geeraerts, 1997; Ullmann, 1957; Bloomfield, 1933; Reisig, 1839). The typology proposed by Blank (1997) has become an important reference in computational linguistics. Cassotti et al. (2024) are the first to employ Blank's typology and data as a testing ground for definition-generation models as tools for semantic change detection. This line of work, building on Periti et al. (2024), continues a research agenda that leverages synchronic lexical and semantic relations, for which we have high-quality and sufficient data documented in lexical databases and lexicographic dictionaries, for the detection of semantic change in historical corpora. Using this approach, Periti et al. (2024) investigated how embedding similarities vary for different change types. Cassotti et al. (2024) extended this by using synchronic pairs of sense definitions to develop a new model to discriminate between types of change.

There are also solid grounds to hypothesize that connotational meaning might help capture processes of semantic change. Such links were noted early on in historical semantics (Van Ginneken, 1911; Sperber, 1914), for instance, on the role of metaphorical expression to attenuate negative emotional value, and have received renewed attention recently (Xu et al., 2021; Fugikawa et al., 2023; Goworek and Dubossarsky, 2024). In addition, change in connotational meaning can occur by itself, as the literature has documented many cases of so-called *amelioration* (*nice* going from 'foolish' to 'agreeable') and *pejoration* (Geeraerts, 2010).

On computational evaluation, current annotated benchmarks include the synchronic, definition- and type-based *LSC Cause-Type-Definitions Benchmark* (Cassotti et al., 2024) and the binary, word-sense-based *TempoWIC*, where semantic change is labeled by comparing the similarity between usages at different time points (Loureiro et al., 2022).

7 Conclusion

The present study introduced the first sense-level benchmark for modeling semantic relations between word senses, uniting denotational and connotational aspects of meanings. It distinguishes denotational relations – *generalization, specialization, metaphor, metonymy* and *homonymy* – as well as two connotational dimensions – *valence* and *arousal*. The expert annotation by historical linguists, for the denotational types, and psychology scholars, for the connotational dimensions, totals 300 pairs of senses for the denotational types (60 for each of the five relations), unrivaled in size by any existing type annotated resource, and 952 sense pairs annotated for each of the connotational dimensions. The benchmark will be publicy released.

Using the benchmark, we evaluate LLMs from three different families, namely GPT-4o, Llama 3.1, and DeepSeek, in a zero-shot setting. For detecting relational types, we fine-tuned Llama and RoBERTa models on existing resources and, due to available data, we also trained our models for the (auto-) antonymy relation. GPT-4o best aligns with human judgments on connotational dimensions, while the significantly smaller RoBERTa model excels at classifying denotational relations.

We conducted the first large-scale study of these relations and find (unsurprisingly) that the auto-antonyms are the class that exhibit the largest amount of *arousal* and *valence* change, followed closely by generalizations and specializations. In general, more sense pairs exhibit *arousal* changes compared to *valence* changes, and very few experience large changes in both dimensions.

Having systems that can classify both denotational and connotational change holds implications across scientific disciplines, from historical linguistics and NLP to the cognitive and social sciences. Extending this benchmark with temporal information will allow it to serve as evaluation data for diachronic studies of meaning emergence, enabling analysis that spans all phases of semantic change. Nevertheless, the confusion matrix indicates that even fine-tuned models exhibit notable confusion for certain relational classes, leaving substantial room for improvement - a direction we aim to pursue in a future shared task. Overall, this work lays the foundation for, and provides preliminary evidence toward, a unified model that integrates both the dimensions and types of semantic change.

8 Limitations

We identify three main limitations for this study, related to the nature of the benchmark, the difficulty of working with lexical types, and the computational challenges of working with LLMs.

The creation of benchmarks, especially semantic ones, always involves a trade-off between quality and coverage. For the present study, we involved experts in their specific domain (i.e., historical linguistics and psychology), to ensure the reliability of the annotations of sense relations, as opposed to layman annotators (c.f., Schlechtweg et al., 2021). Although this choice increases the probability of high-quality annotation, it also necessarily limits the amount of annotations and consequently the possibility of providing multiple annotations per instance. Validity checks were implemented to counterbalance uncertainty, but scaling-up this expert type of sense annotations remains a challenge considering its resource-intensive nature. The present study repeatedly mentions that the use of synchronic lexical resources (e.g., Wordnet, the ANW-dictionary) are good proxies and necessary stepping stones for the modeling and identification of semantic change in actual historical materials. The most appropriate level for modeling these phenomena is therefore at the level of lexical tokens: semantic change involves the frequency change of senses, and such frequencies can only be calculated after determining the sense of an individual corpus occurrence of a word. This paper is therefore part of a larger project that is now moving toward such concrete humanities applications on historical corpora.

LLMs remain opaque. They have limited word sense disambiguation capabilities as, while they perform well in zero-shot settings, they do not surpass state-of-the-art methods, and fine-tuned models with a moderate number of parameters continue to outperform all other models (Basile et al., 2025). Furthermore, LLM performance can vary significantly across tasks - diverging from human ground truth even when prompt-tuned (Pangakis and Wolken, 2024). In the context of this study, it is unclear whether closed-source models were trained on resources that overlap with our datasets, an issue that may contribute to confusion predicting certain classes. Conversely, these models were likely trained on studies introducing normed affective datasets, which could, speculatively, benefit performance on connotational dimensions.

References

Naomi Baes, Nick Haslam, and Ekaterina Vylomova. 2024. A Multidimensional Framework for Evaluating Lexical Semantic Change with Social Science Applications. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1390–1415, Bangkok, Thailand. Association for Computational Linguistics.

Naomi Baes, Raphael Merx, Nick Haslam, Ekaterina Vylomova, and Haim Dubossarsky. 2025. LSC-eval: A general framework to evaluate methods for assessing dimensions of lexical semantic change using LLM-generated synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10905–10939, Vienna, Austria. Association for Computational Linguistics.

Pierpaolo Basile, Lucia Siciliani, Elio Musacchio, and Giovanni Semeraro. 2025. Exploring the word sense disambiguation capabilities of large language models. *arXiv preprint*, arXiv:2503.08662.

Andreas Blank. 1997. Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen. *Max Niemeyer Verlag*.

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

Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaite, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. 2025. Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific reports*, 15(1):11477.

Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the collaborative interlingual index. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 50–57, Bucharest, Romania. Global Wordnet Association.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Gainesville, FL.

Thomas Brochhagen, Gemma Boleda, Eleonora Gualdoni, and Yang Xu. 2023. From language development to language evolution: A unified view of human lexical creativity. *Science*, 381(6656):431–436.

Sven Buechel and Udo Hahn. 2022. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996*.

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.

Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2024a. Semantic Change Characterization with LLMs using Rhetorics. *Preprint*, arXiv:2407.16624.

Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2024b. Survey in Characterization of Semantic Change. *Preprint*, arXiv:2402.19088.

Dutch Language Institute. n.d. Algemeen nederlands woordenboek (anw). Online service.

Olivia Fugikawa, Oliver Hayman, Raymond Liu, Lei Yu, Thomas Brochhagen, and Yang Xu. 2023. A computational analysis of crosslinguistic regularity in semantic change. *Frontiers in Communication*, Volume 8 - 2023.

Dirk Geeraerts. 1997. *Diachronic Prototype Semantics: A Contribution to Historical Lexicology*. Oxford University Press.

Dirk Geeraerts. 2010. *Theories of Lexical Semantics*. Oxford University Press.

Dirk Geeraerts. 2020. Semantic Change: "What The Smurf?". *The Wiley Blackwell Companion to Semantics*, pages 1–24.

Roksana Goworek and Haim Dubossarsky. 2024. Toward Sentiment Aware Semantic Change Analysis. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 350–357, St. Julian's, Malta. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition, chapter Lexicons for Sentiment, Affect, and Connotation. Online manuscript released August 24, 2025.

Temuulen Khishigsuren, Gábor Bella, Thomas Brochhagen, Daariimaa Marav, Fausto Giunchiglia, and Khuyagbaatar Batsuren. 2022. How universal is metonymy? results from a large-scale multilingual analysis. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 96–98, Seattle, Washington. Association for Computational Linguistics.

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

Gonzalo Martínez, Juan Diego Molero, Sandra González, Javier Conde, Marc Brysbaert, and Pedro Reviriego. 2024. Using large language models to estimate features of multi-word expressions: Concreteness, valence, arousal. *Behavior Research Methods*, 57(1):5.

Rowan Hall Maudslay, Simone Teufel, Francis Bond, and James Pustejovsky. 2024. ChainNet: Structured Metaphor and Metonymy in WordNet. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2984–2996, Torino, Italia. ELRA and ICCL.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Charles Egerton Osgood, William H May, and Murray S Miron. 1975. *Cross-Cultural Universals of Affective Meaning*. University of Illionois Press.

Nicholas Pangakis and Samuel Wolken. 2024. Keeping humans in the loop: Human-centered automated annotation with generative ai. *arXiv preprint arXiv:2409.09467*.

Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024. Analyzing Semantic Change through Lexical Replacements. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4495–4510, Bangkok, Thailand. Association for Computational Linguistics.

Karl Reisig. 1839. *Vorlesungen über lateinische Sprachwissenschaft*. Lehnhold.

James A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172.

Dominik Schlechtweg. 2023. Human and computational measurement of lexical semantic change.

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, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hans Sperber. 1914. *Über den Affekt als Ursache der Sprachveränderung*. Niemeyer.

Stephen Ullmann. 1957. *The Principles of Semantics*. Basil Blackwell, Oxford.

Jac Van Ginneken. 1911. Het gevoel in taal en woordkunst 1. *Leuvensche Bijdragen*, 9:265–356.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Xiuwen Wu, Hao Wang, Zhiang Yan, Xiaohan Tang, Pengfei Xu, Wai-Ting Siok, Ping Li, Jia-Hong Gao, Bingjiang Lyu, and Lang Qin. 2025. Ai shares emotion with humans across languages and cultures. *arXiv preprint arXiv:2506.13978*.

Aotao Xu, Jennifer E. Stellar, and Yang Xu. 2021. Evolution of emotion semantics. *Cognition*, 217:104875.

George Kingley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, Cambridge, Massachusetts.

A Prompt Design

A.1 Dimensions

To ensure clarity of constructs in the LLM prompts for each affective dimension, we provided definitions and anchors that combined reliable existing measures from Warriner norms (Warriner et al., 2013) (psycholinguistic approach) and NRC-VAD (Mohammad, 2018) (crowd-sourced annotations approach), covering both ends of the scale. Tasks were formulated as relative difference between sense pairs to avoid order bias (i.e., where annotators are influenced by showing a particular sense first when comparing it to another sense), and the final prompt was designed to be calibrated with anchor terms so both humans and LLMs apply 'consistent' thresholds when annotating.

Recent work shows that LLMs can replicate human analysis across affective tasks including sentiment leaning, emotional intensity (Bojić et al., 2025), and prediction of valence and arousal at the word- and multi-word level (Martínez et al., 2024). LLM-derived emotion spaces have also been demonstrated as being underpinned by *valence* and *arousal* (Wu et al., 2025), indicating that LLMs have the capability to annotate these affective connotational dimensions.

A.1.1 Prompt Optimization

We tested whether prompt wording influenced performance. *Arousal* was chosen as the affective dimension to experiment with as preliminary results indicated that it was the most difficult to annotate. First, we compared the baseline "intensity" framing with alternative *arousal*-based prompts. Results informed our decision to focus on the well-studied and reliable constructs of Valence and Arousal. Experiments were run on September 2025.

Results show that prompt wording strongly affected performance. The baseline intensity framing yielded the weakest correlations, whereas reframing the task in terms of *arousal* — a comparatively reliable measure (Warriner et al., 2013; Mohammad, 2018) — improved alignment with human ratings. In particular, the arousal prompt with anchors drawn from existing scales (Warriner et al., 2013; Mohammad, 2018) achieved the strongest performance ($\rho = .80$ for definition, $\rho = .69$ for sentence). Overall, prompting the model to annotate *arousal* (a construct with a substantial psychological literature) and supplying anchors, as one would for human participants, produced the most reliable results, as demonstrated in Table 5.

Prompt	Definition	Sentence
Baseline (intensity)	.46***	.42***
Arousal v1 (no anchors)	.61***	.67***
Arousal v2 (with anchors)	.80***	.69***

Note. Values are Spearman's ρ. *** $p < .001$.

Table 5: Zero-shot correlations (ρ) between human and LLM judgments under different prompt formulations.

Baseline (Intensity) — Single Definition

Prompt introduction. In psychology research, 'Intensity' is defined as "the degree to which a word's meaning changes to acquire more or less emotionally charged (i.e., strong, potent, high-arousal) connotations." This task focuses on the term {target_word}.
Task. You will be given one sense definition of {target_word}. Your job is to assess how emotionally charged the sense of the target word feels.
Instructions. Classify the intensity of the target word's meaning as either low, neutral, or high.
Response format (return exactly one label):

- LOW – The word is used in a low-intensity, emotionally mild or muted way.
- NEUTRAL – The word is used in a context that

conveys no strong emotional charge; standard or baseline usage.

- HIGH – The word is used in a way that conveys strong emotional force or heightened arousal.

Sense definition: {definition}

Arousal v1 (No Anchors) — Single Definition

Definition. 'Arousal' is the degree of energy, stimulation, or alertness conveyed by the connotations of a word's meaning (Warriner et al., 2013; NRC-VAD, Mohammad, 2018).
Task. You will be given a dictionary-style sense definition of {target_word}. Rate the typical emotional arousal conveyed by this sense. Focus only on the sense definition, not usage context.
Response format (return exactly one label):

- LOW – relaxed, calm, sluggish, dull, sleepy, passive.
- NEUTRAL – steady, ordinary in energy.
- HIGH – stimulated, excited, frenzied, jittery, alert, active.

Sense definition: {definition}

Arousal v2 (With Anchors) — Single Definition

Definition. 'Arousal' is the degree of energy, stimulation, or alertness inherent in a word's meaning, ranging from excited/active to calm/passive (Warriner et al., 2013; NRC-VAD, Mohammad, 2018).
Task. You will be given a dictionary-style sense definition of {target_word}. Rate the emotional arousal conveyed by this sense, using the definition to guide your decision. Focus on the level of energy or activation that the word inherently conveys, ignoring any external context (e.g., example sentences). Use the descriptors below as anchors.
Response format (choose one):

- NEUTRAL – The sense conveys steadiness or ordinary energy, lacking strong arousal in either direction.
- LOW – The sense conveys low energy or arousal (e.g., calmness, dullness, passiveness, relaxation, sleepiness, sluggishness, lack of arousal).
- HIGH – The sense conveys high energy or arousal (e.g., activeness, alertness, excitement, frenzy, jitteriness, stimulation, wakefulness).
- CANNOT DECIDE – If unclear or ambiguous.

Sense definition: {definition}

B Human Pairwise Sense Comparisons

From the 500 sense pairs, we additionally annotated a subset of 50 sense pairs for differences, and used this as an initial test set for model evaluation for the pair of definitions and sentences in the difference conditions. Because results were poor (see Appendices B.1, B.2, B.3), we proceeded with using the Instance annotations for Difference conditions after computationally deriving difference scores. Overall, the unreliable scores signal that the pairwise sense comparison task is difficult for humans. Reasons may include the fact that (1) the lack of direction of difference leads to challenges in interpreting the results; (2) two people can agree that the difference between sense pairs is big but disagree about the direction of that difference; (3) what counts as a large/small difference is subjective and can amplify disagreement.

B.1 Valence

For the 50 sense-pairs rated for their magnitudes of difference in valence, inter-annotator reliability was low to moderate. The correlation between annotators' ordinal ratings was small and nonsignificant ($\rho = .21$, $p = .152$; n = 50). Agreement was 60%, with an unweighted $\kappa = .37$, reflecting fair-to-moderate chance-corrected agreement. Weighted κ values (linear =.30, quadratic =.22) were low, suggesting that disagreements often involved ratings that were more than one category apart (e.g., "BIG" vs. "SMALL"). As illustrated in Figure 3, NB occasionally assigned higher magnitudes of difference ("BIG") where NH tended to assign lower ones ("SMALL"), indicating systematic differences rather than random variation.

B.2 Arousal

For arousal, inter-annotator reliability was very low. The correlation between annotators' ordinal judgments was negative and significant ($\rho = -.34$, $p = .016$; n = 50), indicating systematic disagreement in rank-ordering. Raw agreement was modest (56%), and the unweighted κ of .29 reflected only fair agreement beyond chance. Weighted κ values were very low (linear κ = .02) or negative (quadratic κ = –.32), showing that reliability dropped sharply when more distant mismatches were penalized. This pattern suggests that annotators often made opposing judgments (e.g., one rating "SMALL", the other rating "BIG") rather than only adjacent categories (e.g., "SMALL" vs. "SAME"). As shown in Figure 4, annotators agreed most consistently on the SAME category.

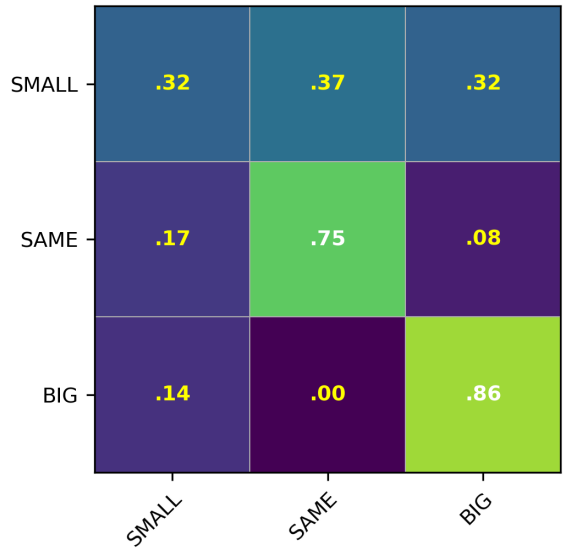

Figure 3: Confusion matrix (row-normalized) comparing NB and NH valence annotations on 50 sense-pair comparisons. Rows = NB; Columns = NH. SMALL = small difference in valence; BIG = big difference in valence.

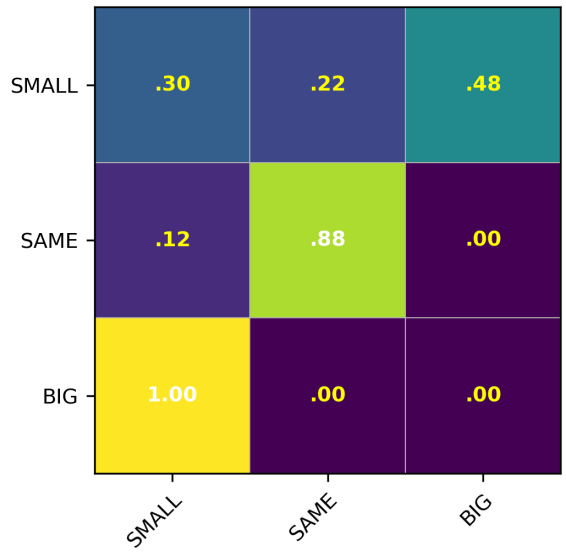

Figure 4: Confusion matrix (row-normalized) comparing NB and NH arousal annotations on 50 sense-pair comparisons. Rows = NB; Columns = NH. SMALL = small difference in arousal; BIG = big difference in arousal.

B.3 Relatedness

Semantic relatedness quantifies the degree to which two uses of a word share meaning. We ran initial experiments to test whether an existing scale could be reliably adapted for this task. First, we adapted the DURel framework (Schlechtweg et al., 2018), which rates relatedness between word usages from 4 (Identical) to 1 (Unrelated), to focus on senses rather than usages and to range from 1-3, dropping the Identical rating, as sense pairs cannot be identical. Unrelated (1) sense meanings correspond to homonymy, where senses diverge entirely; Distantly related (2) senses capture polysemy, frequently realized through metaphorical extensions; and Closely related (3) senses correspond to context variance, often including subtle sense differences or metonymic extensions. The final scale evaluates the degree of semantic relatedness between 952 pairs of senses (247 unique words) on this adapted scale.

The two annotators showed low to moderate reliability. The correlation between their ordinal ratings was positive and significant ($\rho = .47$, $p < .0001$; $n = 952$), indicating some consistency in rank ordering but notable disagreements in categorical assignments. Overall, agreement was 45%, with an unweighted $\kappa = .21$ and weighted values (linear = .29, quadratic = .37) indicating fair chance-corrected agreement. As shown in Figure 5, annotators agreed most consistently on "UNRELATED" judgments (85% of cases) but disagreed more often for "DISTANTLY RELATED" and "CLOSELY RELATED" pairs, which were frequently confused with one another. Consequently, this dimension was excluded from the main analyses due to limited inter-rater reliability, and insufficient evidence to show that it can be extended/applied to the definitional level of meaning.

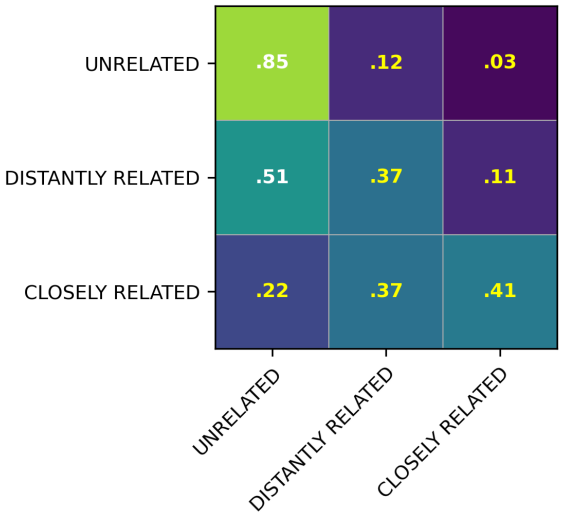

Figure 5: Confusion matrix (row-normalized) comparing NB and NH Relatedness annotations on 50 sense-pair comparisons. Rows = NB, Columns = NH.

C Dimensions Dataset Construction

Gold-standard datasets were constructed for two affective dimensions: Valence and Arousal, to evaluate how well systems, given pairs of definitions and sentences containing the same word, predict the magnitude of difference between sense pairs. Because Arousal has been demonstrated to be more difficult to annotate at the word level, we developed a dedicated seed set based on it.

To identify Arousal seeds, all WordNet lemmas and sense keys were matched to arousal ratings from the NRC Valence, Arousal, and Dominance Lexicon (Mohammad, 2018). This lexicon provides reliable human ratings (split-half reliability = .90) for over 20,000 words, covering substantially more vocabulary (> 40%) than comparable resources. Sense keys were filtered into high (0.82–0.99), neutral (0.47–0.49), and low (0.046–0.25) arousal bands, ensuring at least 250 words per band and starting the range from the median of the categories. Candidates were then separated into monosemous and polysemous cases, and linked to their glosses and examples. For polysemous items, manual annotation was required to align sense-level usage with word-level arousal ratings. The top 100 terms per category (high, neutral, low) were selected and reviewed, with WordNet glosses guiding adjustments where sense-specific intensity diverged from general word ratings. Of 1,905 senses annotated, 141 (71 unique words) were reclassified, mainly from high to neutral. For example, *violate* had six senses, and its use in the sense of 'destroy' contains higher semantic intensity than its use in the sense of 'act in disregard of laws, rules, contracts, or promises'. Also, some words with contextual or cultural shifts (e.g., metaphorical usage or outdated societal norms) were re-annotated accordingly. Two native English speakers (female, male) annotated in a blinded setting, with disagreements resolved through discussion. The final seed set contained balanced monosemous and polysemous items across categories.

This balanced gold-standard Arousal seed set was then used to train RoBERTa[3] to predict the arousal levels (low, neutral, high) of other senses based on their WordNet sense glosses. This synthetic dataset of 23,997 senses was then used to select 500 sense pairs with the largest differences in arousal (after pairing senses to get their difference classifications as "unchanged", "increase", or "decrease"). These were subsequently annotated for their degrees of difference on all three dimensions, placing an emphasis on sense definitions but also considering usage examples.

We assessed inter-annotator reliability between experts using Spearman's ρ and Cohen's κ, which capture complementary aspects of agreement: Spearman's ρ evaluates rank-order consistency across ordinal categories, while Cohen's κ provides a chance-corrected measure of categorical agreement, with weighted variants distinguishing between near and extreme disagreements. Across the 662 annotated sense pairs (235 unique words), the annotators produced broadly similar category distributions. Reliability was moderate for arousal ($\rho = .595$, $p < .0001$; $\kappa = .520$–$.576$ weighted), with most disagreements involving adjacent categories, and higher for polarity ($\rho = .831$, $p < .0001$; $\kappa = .741$–$.830$), reflecting substantial to near-perfect agreement.

D Types Dataset

The statistics of the dataset are reported in Table 6.

E Types: Generation parameters, Model training, and Hyper-parameters

Finetuning RoBERTa-Large Definitions were concatenated, and a linear layer followed by a softmax activation was added on top to predict the label. A linear learning rate schedule was used, with the number of warm-up steps set to 10% of the total training steps. The model was trained for up to 10 epochs with a batch size of 8. The learning rate was selected from the set {1e-4, 2e-4, 1e-5, 2e-5, 1e-6}. The best model was chosen based on the weighted F1 score on the development set, with the optimal learning rate being **1e-6**.

Finetuning Llama 3.1 8B Definitions were concatenated using a special token `<|s|>`, and the symbol `<|t|>` was appended before the label. The model was trained to predict the label following `<|t|>` using causal language modeling. Labels (e.g., *generalization*, *specialization*, etc.) were also encoded as special tokens such as `<|generalization|>`. During inference, the predicted label was determined by selecting the token with the highest logit score immediately after `<|t|>`, restricted to the label-specific tokens and ignoring other entries in the vocabulary.

[3]FacebookAI/roberta-base: https://huggingface.co/FacebookAI/roberta-base

Class	Train			Dev			Test		
	ChainNet	UniMet	WordNet	ChainNet	UniMet	WordNet	ChainNet	UniMet	WordNet
generalization	0	0	30000	0	0	500	0	0	500
homonymy	1269	0	0	500	0	0	500	0	0
metaphor	2666	0	0	500	0	0	500	0	0
metonym	2503	2101	0	254	246	0	279	221	0
specialization	0	0	30000	0	0	500	0	0	500

Table 6: Statistics for train, dev, and test sets across data sources.

We used QLoRA with `nf4` quantization, LoRA $\alpha = 16$, LoRA rank $= 8$, and LoRA dropout $= 0.1$. A linear learning rate schedule was applied, with 10% of total steps used for warm-up. The model was trained for up to 10 epochs with a batch size of 8. The learning rate was searched over {1e-4, 2e-4, 1e-5, 2e-5, 1e-6}. The best model was selected based on the weighted F1 score on the development set, with the best learning rate being **1e-5**.

Zero-Shot Setting For **DeepSeek-V3.2-Exp** and **GPT-4o**, we used the official APIs with default parameters for chat completion. For **Llama 3.1 8B Instruct** and **Llama 3.1 70B Instruct**, predictions were generated using greedy search.

F Sign flips

In addition to what is discussed in the main paper, we also studied the proportion of types that experienced a sign flip. Antonymy, homonymy, and metaphor pairs displayed the highest valence-flip rates (around 0.05), indicating frequent shifts between opposite emotional polarities, consistent with their oppositional or cross-domain nature. Metonymic and taxonomical pairs exhibited much lower valence-flip rates (approximately 0.01–0.02), reflecting affective stability within closely related or hierarchical senses. For *arousal*, the largest flip proportion occurred in *antonymy* (about 0.05), followed by *taxonomical* and *metonymic* relations (around 0.03), while *metaphor* showed a lower arousal-flip rate (about 0.02) and *homonymy* was near zero, likely reflecting its small sample size. Overall, relations involving cross-domain mappings (metaphor) or oppositional meaning (antonymy, homonymy) tended to produce emotional polarity reversals, whereas relations based on association or taxonomy preserved affective orientation.

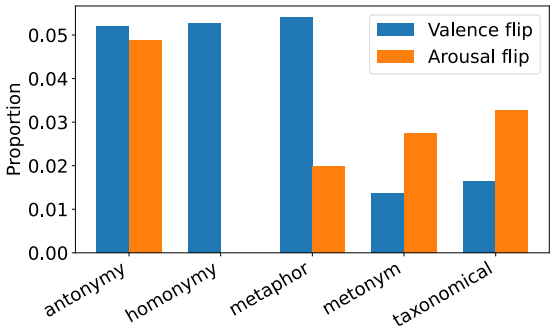

Figure 6: Sign-flip rates by predicted type.

G WordNet Hierarchy experiment

We use the denotational types to investigate how different types of semantic relations reflect structural organization in the lexical hierarchy, we examined the correspondence between relation labels and their topological distances in WordNet. The goal is to determine whether conceptual relations such as metaphor or metonymy, which involve meaning extension across domains, are also manifested as increased separation within the taxonomic graph.

We restricted the analysis to **nouns and verbs**, the only WordNet parts of speech organized into a hypernym–hyponym taxonomy. For each pair of related senses, the corresponding WordNet synsets were retrieved and their **shortest-path distance** within the taxonomy. This measure returns the length of the shortest sequence of hypernym or hyponym links connecting the two synsets, thus quantifying their hierarchical separation. Pairs lacking a defined path (i.e., belonging to disconnected subgraphs) were excluded. The resulting dataset included pairs grouped by their predicted semantic relation label: 201 *taxonomical*, 4 *antonymy*, 155 *metaphor*, 104 *metonym*, and 6 *homonymy* pairs.

Analysis of WordNet path distances across semantic relation types revealed systematic differences in hierarchical proximity (Figure 7). *Taxo-*

nomical relations exhibited the shortest mean distance ($M = 7.29$), reflecting senses that share the same conceptual branch through direct hypernym–hyponym links. *Antonymy* pairs ($M = 7.75$, $SD = 6.75$) showed similar proximity but greater variability, as opposites often occupy parallel subtrees within a shared domain.

By contrast, *metaphorical* and *metonymic* relations showed substantially larger distances. *Metaphor* pairs ($M = 9.81$) typically involve a **domain shift**, where meaning extends from a concrete source to an abstract or functionally distinct target domain (e.g., *grasp* in the physical vs. cognitive sense). This cross-domain mapping separates the senses into distant branches of the lexical hierarchy. *Metonymic* pairs ($M = 10.64$) connect **conceptually contiguous but distinct areas** within a domain, such as part–whole or container–content relations (e.g., *crown* $\rightarrow$ *monarch*), producing moderate hierarchical separation.

Finally, *homonymy* pairs ($M = 12.67$) displayed the largest mean distance, consistent with their etymological independence and lack of conceptual overlap. The overall gradient, i.e. **taxonomical < antonymy < metaphor < metonym < homonymy**, demonstrates a continuum of conceptual relatedness, such that as taxonomic distance increases, the cognitive and semantic association between senses diminishes

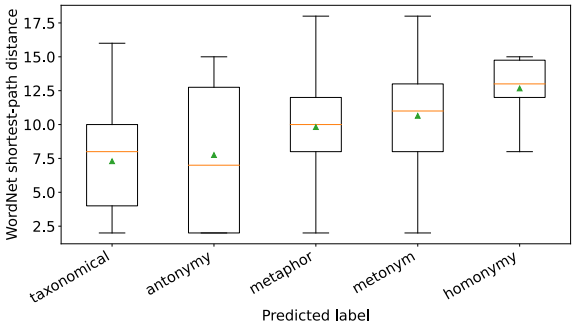

Figure 7: Distance on WordNet hierarchy by predicted type.