A General Framework to Evaluate Methods for Assessing Dimensions of Lexical Semantic Change Using LLM-Generated Synthetic Data

Anonymous ACL submission

Abstract

Lexical Semantic Change (LSC) offers insights into cultural and social dynamics. Yet, the validity of methods for measuring kinds of LSC has yet to be established due to the absence of historical benchmark datasets. To address this gap, we develop a novel three-stage evaluation framework that involves: 1) creating a scalable, domain-general methodology for generating synthetic datasets that simulate theory-driven LSC across time, leveraging In-Context Learning and a lexical database; 2) using these datasets to evaluate the effectiveness of various methods; and 3) assessing their suitability for specific dimensions and domains. We apply this framework to simulate changes across key dimensions of LSC (SIB: Sentiment, Intensity, and Breadth) using examples from psychology, and evaluate the sensitivity of selected methods to detect these artificially induced changes. Our findings support the utility of the synthetic data approach, validate the efficacy of tailored methods for detecting synthetic changes in SIB, and reveal that a state-of-the-art LSC model faces challenges in detecting affective dimensions of LSC. This framework provides a valuable tool for dimension- and domain-specific benchmarking and evaluation of LSC methods, with particular benefits for the social sciences.

1 Introduction

Lexical Semantic Change (LSC) provides a unique window into cultural dynamics by revealing how language evolution reflects social changes. Recently developed state-of-the-art (SOTA) computational methods have expanded our ability to classify established types of LSC, such as generalization and specialization (Cassotti et al., 2024a). Efforts have also been directed towards developing methods for measuring newly proposed dimensions of LSC (Baes et al., 2024; de Sá et al., 2024). Nevertheless, the field faces challenges in validating these methods. A major obstacle is the absence of historical benchmark datasets, which restricts the standardization and fair comparison of metrics. Additionally, there is a pressing need for fine-grained evaluation methods that save time and resources.

To address these challenges, the present study introduces a three-stage evaluation framework. It: 1) develops a scalable, domain-general methodology for generating high-quality synthetic sentences that leverage In-Context Learning (ICL) and a lexical database to simulate changes in kinds of LSC; 2) uses these newly constructed historical datasets to evaluate the relative effectiveness of computational approaches; and 3) identifies the more suitable method for specific dimensions and domains. This framework is applied to assess the sensitivity of various methods to detect synthetic change in major LSC dimensions–Sentiment, Intensity, and Breadth (SIB; Baes et al. 2024)–using examples drawn from psychology. Our findings confirm the validity of theory-driven changes using synthetic SIB datasets and emphasize the need to tailor methods to particular dimensions, as the SOTA LSC model was found to be ineffective at detecting affective dimensions. This framework provides an efficient and scalable solution for dimension- and domain-specific benchmarking and evaluation of LSC methods. While this innovation is generally applicable, it is particularly beneficial for the social sciences and humanities, where customized methods are essential for analyzing complex constructs.

2 Related Work

2.1 Theoretical Background

Linguists have long debated taxonomies of LSC (Bloomfield, 1933; Blank, 1999), defined as innovations which change the lexical meaning of a form (Bloomfield, 1933). A growing body of work has identified ways to detect changes in the meanings of words and quantify the extent of these changes using a variety of computational approaches (Ku-

tuzov et al., 2018; Tahmasebi et al., 2018; Tang, 2018; Cassotti et al., 2024b; Periti and Montanelli, 2024a; Kiyama et al., 2025).

Recent years have seen the development of theoretical frameworks that propose multiple dimensions of LSC. Baes et al. (2024) introduced a three-dimensional framework that maps LSC along axes of SIB, reflecting a word's acquisition of more positive or negative connotations (Sentiment), more or less emotionally charged or potent connotations (Intensity), and the expansion or contraction of its semantic range (Breadth). It draws on linguistic (Geeraerts, 2010) and psychological (Haslam, 2016) theories, and provides methodological tools to estimate SIB across time. In parallel, de Sá et al. (2024) proposed a framework that clusters LSC into three dimensions using graph structures: Orientation (shifts towards more pejorative or ameliorated senses), Relation (changes towards metaphoric or metonymic usage), and Dimension (variations between abstract/general and specific/narrow meanings). While de Sá et al. (2024) surveyed statistical methods for representing word meaning (word frequency, topic modeling, and graph structures) on dimensions, they did not demonstrate their usage.

Both frameworks contain dimensions of evaluation (Sentiment and Orientation) and semantic range (Breadth and Dimension). Baes et al.'s (2024) inclusion of Intensity reflects a greater emphasis on changes in the emotional connotations of words. Sentiment and Intensity resemble the two primary dimensions of human emotion, Valence and Arousal (Russell, 2003), and two primary dimensions of connotational meaning, Evaluation (e.g., "good/bad") and Potency (e.g., "strong/weak") (Osgood et al., 1975), which have been demonstrated to have cross-cultural validity.

2.2 Evaluation

Despite substantial progress in developing benchmarks (Tahmasebi and Risse, 2017) and evaluation strategies (Kutuzov et al., 2018), the field still lacks standardized datasets that evaluate multiple dimensions of LSC across time. Current annotated benchmarks, such as the synchronic, definition- and type-based *LSC Cause-Type-Definitions Benchmark* (Cassotti et al., 2024a) and the binary, word-sense-based *TempoWIC*, where LSC is labeled by comparing the sameness or difference of meanings between two sense usages (Loureiro et al., 2022), address different aspects of semantic change.

The first human-annotated dataset of LSC in multiple languages (English, German, Latin, Swedish; Schlechtweg et al., 2020) represented substantial progress in indicating the presence and degree of LSC, but omitted information about kinds of change. Creating expert-annotated datasets of LSC is costly and time-intensive. Recognizing this gap, Dubossarsky et al. (2019) introduced a method to artificially induce semantic change in controlled testing environments, allowing for precise testing of how well models capture these shifts.

Recent developments in generative artificial intelligence highlight the potential of pre-trained LLMs to adapt to novel tasks at inference time through ICL (Zhou et al., 2023). Few-shot ICL, a paradigm that enables LLMs to learn tasks by analogy given only a few demonstrative examples, helps to incorporate theoretical knowledge without needing to fine-tune its internal parameters (Dong et al., 2024). Instead, ICL uses context from the model's prompt to adapt the LLM to downstream tasks (Radford et al., 2019; Brown et al., 2020; Liu et al., 2024). de Sá et al. (2024) demonstrated the utility of few-shot ICL, employing Chain-of-Thought and rhetorical devices, to annotate LSC dimensions, but their strategy focuses on multi-class classification of change between two sense usages. ICL offers a promising solution to bridge the absence of standardized approaches (Hengchen et al., 2021) for assessing the effectiveness of different methods to measure dimensions of LSC.

2.3 The Present Study

The present study aims to develop an evaluation framework that: (1) creates a scalable, domain-general methodology for constructing high-quality LLM-generated datasets labeling changes in LSC dimensions; (2) uses these synthetic datasets to compare the validity of proposed computational approaches; and (3) identifies the suitability of methods for each dimension and domain. We apply this framework to major LSC dimensions defined by Baes et al. (2024) (SIB; see Table 1) on a sample of words drawn from a corpus of academic psychology articles. Key questions include:

1. Can synthetic datasets validate methods to measure dimensions of LSC? We predict that SIB scores will be linearly associated with levels of synthetic change.
2. Which out of a set of LSC detection methods is most sensitive to synthetically induced changes in SIB?

Dimension	Definition	Examples of Rising	Examples of Falling
Sentiment	Relates to the degree to which a word's meaning acquires more positive (*'elevation', 'amelioration'*) or negative (*'degeneration', 'pejoration'*) connotations.	*craftsman*, once associated with manual labor, has now come to convey artistry, skill, and high-quality workmanship. *geek*, from a derogatory term for odd people, to reference someone passionate about a specific field.	*retarded*, originally a neutral term for intellectual disability, has become highly pejorative over time. *awful* has shifted from its original meaning of "awe-inspiring" to its modern usage which indicates something very bad.
Intensity	Relates to the degree to which a word's meaning changes to acquire more (*'meiosis'*) or less (*'hyperbole'*) emotionally charged (i.e., strong, potent, high-arousal) connotations.	*cool* has evolved from describing temperate to expressing strong approval or trendiness. *hilarious*, originally meaning cheerful or amusing in Latin, has come to describe extremely funny things that cause great merriment and laughter.	*love*, evolved from a romantic or platonic attachment to a milder expression of liking (e.g., "I love pizza.") *trauma*, from referencing brain injuries to referring to less severe events (e.g., business loss).
Breadth	Relates to the degree to which a word expands (*'widening', 'generalization'*) or contracts (*'narrowing', 'specialization'*) its semantic range.	*cloud*, initially a meteorological term, broadened its use to reference internet-based data storage. *partner*, originally referring to business co-owners, now also describes a significant other in a romantic or domestic relationship.	*doctor*, once referring to any scholar or teacher, now primarily refers to a medical professional. *meat*, originally referred to any kind of food in Old English ('mete'), but its meaning has narrowed to specifically denote animal flesh as food.

Table 1: Definitions and Examples of Baes et al.'s (2024) Dimensions of Lexical Semantic Change.

3 Method

3.1 Materials

3.1.1 Psychology Corpus

To develop and test the evaluation pipeline on a specific domain, a corpus of psychology article abstracts was sourced (Vylomova et al., 2019). It includes 133,017,962 tokens from 871,337 abstracts (1970-2019) from E-Research and PubMed databases, and contains 5,214,227 sentences.[1]

3.1.2 WordNet

Although other ontologies were considered,[2] the English WordNet lexical database 3.0 (Miller, 1992) was chosen for its linguistic coverage and lexical structure. It organizes words into synsets (synonyms with distinct meanings), linking them by semantic relationships (e.g., hypernyms, hyponyms).

3.1.3 Targets

While the evaluation framework is general in its applicability, six terms from psychology—*abuse*, *anxiety*, *depression*, *mental health*, *mental illness*, and *trauma*—are analyzed for semantic change, selected for their empirical and theoretical relevance to shifting word meanings. *Trauma*, *mental health*, and *mental illness* have seen falls in their average valence and semantic expansions (*trauma*: Baes et al., 2023; Haslam et al., 2021; *mental health*, *mental illness*: Baes et al., 2024). There have been changes in the intensity of their meanings, with rises for *mental health* and *mental illness* (Baes et al., 2024), as well as *anxiety* and *depression* (Xiao et al., 2023), and a fall for *trauma* (Baes et al., 2023). Qualitatively, the concept of *abuse* has expanded horizontally to include passive neglect and emotional abuse, beyond its initial physical scope (Haslam, 2016). Each target was sufficiently prevalent in the corpus for robust analysis, with sentence counts at 46,272; 104,486; 115,430; 44,130; 5,808; and 23,187 (see Appendix A for annual counts).

3.2 Evaluation Framework

The general pipeline for the evaluation framework is presented in Figure 1. First, synthetic datasets are constructed to benchmark changes in LSC dimensions using few-shot ICL and a lexical database. GPT-4o (Achiam et al., 2023)[3] is prompted with expert-crafted examples to increase and decrease corpus sentences in affective dimensions across 5-year intervals. This ensures that synthetic sentences are theory-driven, domain-specific and contain temporal features. GPT is used due to its adeptness at few-shot learning, task adaptation with minimal examples (Achiam et al., 2023; Merx et al., 2024) and lack of disciplinary bias (Ziems et al., 2024). Appendix B details statistics for synthetic datasets,[4] which are validated using tools shown to measure SIB in two corpora (Baes et al., 2024).

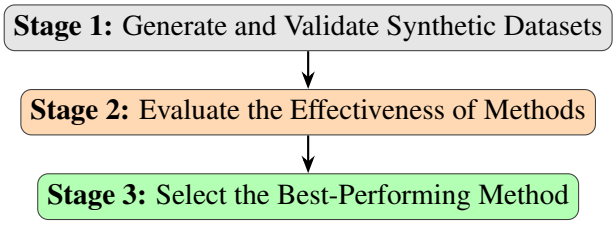

Figure 1: General Stages for the Evaluation Framework.

[1] Sentences were segmented using "en_core_web_sm" (https://spacy.io/models/en); F-score = 91%.

[2] PsycNET, UMLS, DSM-5, ConceptNet

[3] ChatGPT API documentation: https://platform.openai.com/docs/guides/text-generation

[4] Link to Synthetic datasets: [MASKED LINK].

To assess the relative effectiveness of different methods, sentences are sampled from the natural and synthetic corpora using two sampling strategies. Bootstrap sampling draws 50 sentences with replacement from both corpora 100 times (i.e., iterations). Five-year random sampling selects up to 50 sentences from fixed intervals 10 times, ensuring each time period is equally represented. Each iteration forces unique sentence selection while permitting sentence repetition across different rounds to reflect natural language. Control conditions shuffle these sentences to balance authentic:synthetic sentences in each sample, verifying the synthetic effect in the genuine condition and its absence in the control. Following computational linguistics precedents (Dubossarsky et al., 2017, 2019), this approach validates the impact of synthetic interventions by providing a baseline for comparison. Notably, for each strategy, synthetic sentences are injected into natural samples at increasing injection levels (20%, 40%, 60%, 80%, and 100%), as illustrated in Appendix B. Bins are injection levels (for bootstrap) and epoch (for 5-year intervals). This simulates controlled semantic saturation scenarios to assess the sensitivity of methods to semantic variation (increased/decreased SIB) (stage 2) and select the method that detects greater magnitude of change from 0% to 100% injection (stage 3).

3.3 Sentiment and Intensity

3.3.1 Synthetic Sentiment and Intensity

To generate the synthetic Sentiment and Intensity datasets, we employ few-shot ICL with GPT-4o to vary these dimensions. First, neutral sentences from the corpus (detailed in Section 3.1.1) are sampled as outlined in Appendix C. Second, a psychology scholar crafts five (Chen et al., 2023) diverse examples of sentence variations for each target following the task detailed below, which includes construct definitions to generate theory-driven change. For 'scholar-in-the-loop' few-shot demonstrations, see Appendix D (Sentiment) and Appendix E (Intensity). Third, the prompt is refined during pilot tests (10 inputs). Fourth, for each of the neutral sentences (Sentiment: 36,151; Intensity: 39,896), we make one inference call to GPT-4o through the OpenAI API to generate variations of Sentiment (positive/negative) or Intensity (high/low). Fifth, output sentences are manually adjusted (Sentiment: 0.25%; Intensity: 0.01%) due to GPT-4o's failure to retain targets. See Appendix C for the counts of input/output sentences and final prompts (dataset costs for Sentiment: 115 $US; Intensity: 136 $US).

Prompt Outline for Synthetic *Sentiment*

Prompt intro: In psychology research, 'Sentiment' is defined as "a term's acquisition of a more positive or negative connotation." This task focuses on the sentiment of the term `target_word`.
Task: You will be given a sentence containing the term `target_word`. Your goal is to write two new sentences:

1. One where `target_word` has a **more positive connotation**.
2. One where `target_word` has a **more negative connotation**.

Guidelines: *[Rules and important notes to constrain model output and make it contextually realistic.]*

Append few-shot examples: *[One example below.]*

Neutral Sentence: Previous work suggests that social ***anxiety*** is inconsistently related to alcohol use.
Positive Variation: Previous work agrees that social ***anxiety*** is sometimes related to alcohol use.
Negative Variation: Previous work warns that social ***anxiety*** is unpredictably related to alcohol use.

Prompt Outline for Synthetic *Intensity*

Prompt intro: In psychology research, 'Intensity' is defined as "the degree to which a word has emotionally charged (i.e., strong, potent, high-arousal) connotations." This task focuses on the intensity of the term `target_word`.
Task: You will be given a sentence containing the term `target_word`. Your goal is to write two new sentences:

1. One where `target_word` is **less intense**.
2. One where `target_word` is **more intense**.

Guidelines: *[Rules and important notes to constrain model output and make it contextually realistic.]*

Append few-shot examples: *[One example below.]*

Neutral sentence: They tend to be more liberal in their attitudes toward abortion than women in general; however, women who experienced a greater degree of psychic ***trauma*** tended to be more conservative in their attitudes.
Low Variation: They tend to be more accepting in their attitudes toward children than women in general; however, women who experienced mild psychic ***trauma*** tended to be more conservative in their attitudes.
High Variation: They tend to be more extremely callous in their attitudes toward the horrors of abortion than women in general; however, women who suffered a greater degree of violent psychic ***trauma*** tended to be more fearful in their attitudes.

3.3.2 Quantifying Sentiment and Intensity

To measure shifts in a word's connotations from negative to positive Sentiment and from low to high Intensity, we adapt Baes et al.'s (2024) method. Sentences are processed.[5] Collocates (±5 words from the target) within sentences are assigned ordinal valence or arousal scores based on Warriner et al. (2013) norms, ranging from *extremely unhappy* (1: "unhappy", "despaired") to *extremely happy* (9: "happy", "hopeful") for valence, and from *extremely low* (1: "calm", "unaroused") to *extremely high* (9: "agitated", "aroused") for arousal. Valence (V) and arousal (A) indices are calculated as shown in Equation 1:

$$V_{t_j,k}, A_{t_j,k} = \frac{\sum_{i=1}^{n_{j,k}} w_{i,j,k} x_{i,j,k}}{\sum_{i=1}^{n_{j,k}} w_{i,j,k}} \quad (1)$$

where $w_{i,j,k}$ denotes the frequency of each collocate i in iteration k within bin t_j, and $x_{i,j,k}$ denotes its valence or arousal rating at bin t_j within iteration k. Here, $n_{j,k}$ is the number of collocates in iteration k within bin t_j. Scores are weighted by the collocate's frequencies within each iteration and normalized by the total occurrences in that iteration. Scores are averaged across all iterations within each bin, conditioned on whether the Sentiment is positive/negative, or the Intensity is high/low. These indices provide a mean valence or arousal score per iteration in each bin t_j, with higher scores indicating a more positive valence or higher arousal. Scores (1-9) are normalized to range from 0 (extremely unhappy/low arousal) to 1 (extremely happy/high arousal).

While the Intensity dimension is novel and lacks existing comparative models, for Sentiment, we compare the interpretable Valence index against DeBERTa-v3-ABSA, a SOTA classification model in aspect-based sentiment analysis (ABSA). Deberta-v3-base-absa-v1.1[6] identifies sentiment associated with particular aspects of an entity within text (here, the target term). It was initially trained on restaurant and laptop reviews (Cabello and Akujuobi, 2024; Yang et al., 2021, 2023). We adapt it to produce continuous sentiment scores, which reflect the model's confidence in positive sentiment associated with the target term and range from 0 (fully negative) to 1 (fully positive).[7]

3.4 Breadth

3.4.1 Synthetic Breadth

Unlike Sentiment and Intensity, current Breadth measures have no score that assigns a mid-point with which to obtain neutral sentences to vary. Therefore, to simulate semantic breadth, we adapt Dubossarsky et al.'s (2019) replacement strategy, using WordNet 3.0 to expand a target word's usage by incorporating contexts from donor terms, broadening its semantic range without altering its core meaning. Relevant synsets are identified and filtered for psychological relevance using keyword matching[8] and semantic similarity thresholds. Donor terms (co-hyponyms with the target) are filtered using Lin similarity (0.5)[9] and cosine similarity (0.7) with embeddings from BioBERT (Lee et al., 2020), a pre-trained language model for biomedical text mining, to capture context-dependent meanings of synset glosses in 768-dimensional vectors. See Appendix F for the list. The sibling replacement process identifies and replaces sibling terms with the target, shown below. To sample representatively from the sibling list, a round-robin strategy is used, sampling up to 1,500 unique sentences per epoch per injection level to create the final synthetic breadth dataset (0 $US).

Dataset Creation for Synthetic *Breadth*

Replacement Strategy: Randomly sample sentences containing co-hyponyms of the target term from the validated list and replace the **`co-hyponym`** with the **`target`** to be used as a synthetic sentence.

*[One example for **mental_health** below.]*

Donor Context: The 'Angry and Impulsive Child' and 'Abandoned and Abused Child' modes uniquely predicted ***dissociation*** scores.
Synthetic Context: The 'Angry and Impulsive Child' and 'Abandoned and Abused Child' modes uniquely predicted ***mental_health*** scores.

3.4.2 Quantifying Breadth

To estimate the semantic broadening (expansion) or narrowing (contraction) of a word's meaning, we calculate the average cosine distance between sentence-level embeddings of a target term, as in Baes et al. (2024). The SentenceTransformer

[5] Tokenization, lemmatization, stop-word removal using "en_core_web_sm" (https://spacy.io/models/en)

[6] yangheng/deberta-v3-base-absa-v1.1 (184M model params): https://huggingface.co/yangheng/deberta-v3-base-absa-v1.1

[7] The sentiment score is calculated as follows: $0 \times$ negative_prob $+ 0.5 \times$ neutral_prob $+ 1 \times$ positive_prob.

[8] Psychology key terms: "abnormality", "abnormally", "emotional", "feeling", "feelings", "harm", "hurt","mental", "mind", "psychological", "psychology", "psychiatry", "syndrome", "therapy", "treatment".

[9] Information content values from the psychology corpus

model 'all-mpnet-base-v2'[10] is used to generate these embeddings. The Breadth score, B, is derived by averaging the cosine distances, δ, across all unique pairs of sentence embeddings within each iteration, and then averaging these scores across all iterations within each bin, as shown in Equation 2:

$$B_{t_j} = \frac{1}{I_j} \sum_{k=1}^{I_j} \left(\frac{2}{N_k(N_k-1)} \sum_{i=1}^{N_k-1} \sum_{j=i+1}^{N_k} \delta(s_{i,k}^{t_j}, s_{j,k}^{t_j}) \right) \quad (2)$$

Here, $\delta(s_{i,k}^{t_j}, s_{j,k}^{t_j})$ calculates the cosine distance between two sentence embeddings in the same iteration k in bin t_j. N_k is the number of sentence embeddings in iteration k; I_j is the number of iterations in bin t_j. Higher scores indicate greater variation in the target's semantic range. Scores range from 0 (no variation) to 1 (max variation).

We compare the sentence transformer "all-mpnet-base-v2" (MPNet) with Cassotti et al.'s (2023) SOTA word transformer "XL-LEXEME"[11] (XLL). While MPNet generates sentence embeddings through pooling tokens, which dilutes word-specific information, XLL uses a bi-encoder architecture that focuses on word-specific attention,[12] using polysemy as a proxy for meaning divergence during training (WIC; Pilehvar and Camacho-Collados, 2019).

3.5 General Lexical Semantic Change

To quantify general LSC, we use the SOTA LSC score (Cassotti et al., 2023), which calculates the Average Pairwise Cosine Distances (Giulianelli et al., 2020) between sentence embeddings from two time periods. We extend it to compare embeddings from different bins within the same iteration, as shown in equation 3:

$$LSC_i(s_i^{t_0}, s_i^{t_1}) = \frac{1}{N_i^2} \sum_{m=1}^{N_i} \sum_{n=1}^{N_i} \delta(s_{m,i}^{t_0}, s_{n,i}^{t_1}) \quad (3)$$

Here, N_i represents the number of sentence embeddings within each iteration i in each bin. The term $\delta(s_{m,i}^{t_0}, s_{n,i}^{t_1})$ measures the cosine distance between pairs of sentence embeddings from the same iteration i across two different bins t_0 and t_1. Higher LSC scores indicate greater LSC, ranging from 0 (no change) to 1 (maximum change).

[10]Microsoft pretrained network (109M model params) https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[11]XL-LEXEME (~550M model params) https://huggingface.co/pierluigic/xl-lexeme

[12]Only the first occurrence of the target is attended to.

4 Results

Synthetic Change Effects: The hypothesis that scores from Baes et al.'s (2024) SIB tools will be linearly associated with levels of synthetic change is supported, as evidenced by rising or falling trends in SIB scores across all targets and conditions (Figure 2). Mixed linear models demonstrate increases or decreases on SIB scores for every 1-unit increase in synthetic injection level (detailed in Table 2 and Appendix G). SIB scores for the five-year sampling experiments depict similar trends in response to varying injection levels (Appendix H).

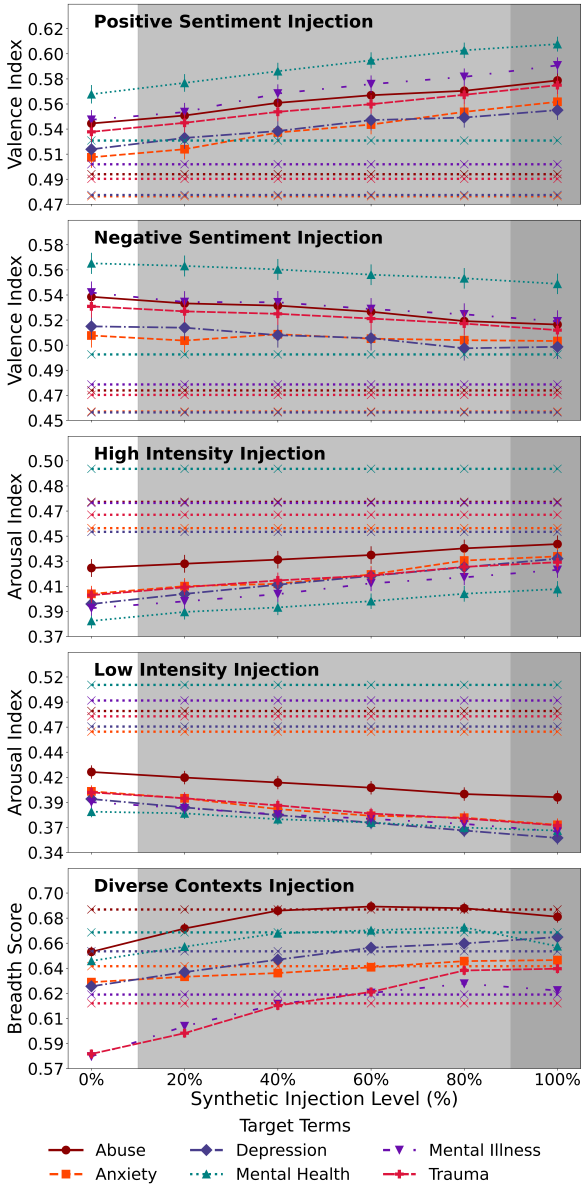

Figure 2: SIB Scores (±SE) by Injection Levels for Experimental and Control Settings (Flat Dotted Lines).

Control Experiments: As illustrated in Figure 2, controlling for synthetic injection level by re-

Score	Valence	Arousal	Breadth
β^+	.003*	.002*	<.0001*
β^-	-.001*	-.002*	N/A

Table 2: Coefficients of Mixed Linear Models Predicting SIB Scores from Injection Levels (Var. Target Intercept)

Note: β^+ and β^- represent the standardized coefficients for conditions (rise/fall), respectively. '*' indicates $p < .0001$, testing the null hypothesis that $\beta = 0$.

analyzing data with shuffled sentences for uniform distribution reveals flat SIB score trends in bootstrapped settings. Appendix H shows that, even with temporal shuffling within time bins, SIB scores in five-year samples tend to converge to a midpoint between natural and synthetic data.

Comparative Method Evaluation: Comparisons of the relative validity of alternative change detection methods yielded mixed results. To determine which method is more sensitive to synthetically induced changes in SIB, we compare their performance on a synthetic change detection task using an evaluation metric specified below.[13]

Percent Relative Change Index

Synthetic Change Detection Task: Detect the magnitude of change in the target word's context when sampling sentences from a natural corpus (0%) and an entirely synthetic corpus (100%).

Percent relative change Δ is defined as:

$$\Delta\% = \frac{X_{100} - X_0}{X_0} \times 100$$

where X_0 represents the measure's score at 0% synthetic injection (natural data) and X_{100} at 100% synthetic injection (fully artificial data).

For Sentiment, Valence index and ABSA's Sentiment score are sensitive to detecting variations in synthetic Sentiment, although the ABSA score outperforms the Valence index 10/12 times. For Intensity, the Arousal index shows sensitivity to detecting variations in synthetic Intensity. For Breadth, XLL outperforms MPNet (4/6 times) on detecting rises in synthetic Breadth using the Breadth score.

Critically, XLL-LSC score is completely insensitive to detecting changes in either Sentiment or

[13]For XLL's LSC Score, Δ is normalized against the intrinsic within-bin variability in both bins of interest:

$$\Delta = \frac{\text{APD}(X_{100}\text{-between-}X_0)}{\max\left[\text{APD}(X_0\text{-within-}X_0), \text{APD}(X_{100}\text{-within-}X_{100})\right]}$$

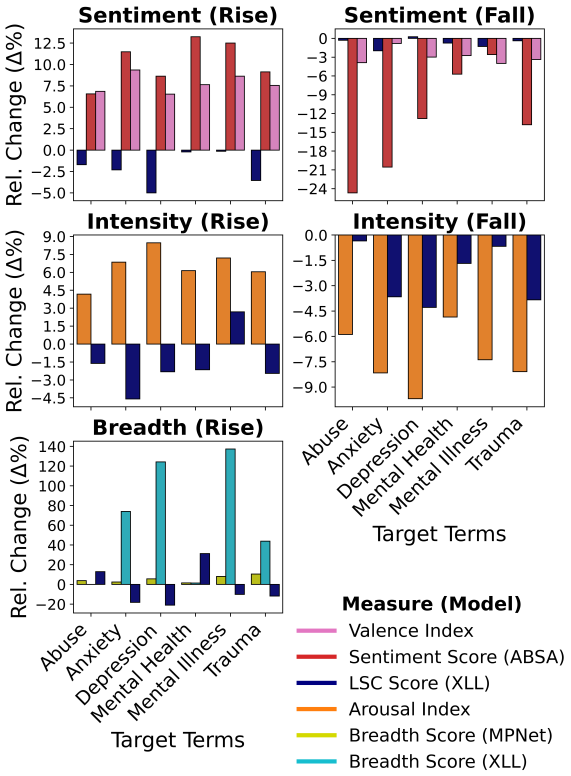

Figure 3: Relative Change ($\Delta\%$) Scores for Models Across Dimensions and Conditions: Bootstrapped.

Intensity. XLL-LSC can only indicate change via positive change values, while negative values indicate that the within-bin variance is greater than the change scores between bins. See Appendix I for between- and within-bin LSC scores across *all* synthetic injection levels. Thus, the negative scores observed in Sentiment and Intensity (except for *Mental Illness*) establish that XLL was unable to detect any change signal in these words. XLL-LSC detects changes in synthetic Breadth for 2/6 terms.

5 Discussion

The present study introduced a three-stage general domain evaluation framework that: 1) creates synthetic datasets featuring 'scholar-in-the-loop' LLM-generated sentences to simulate various kinds of LSC; 2) leverages these datasets to assess the sensitivity of computational approaches to synthetic changes; and 3) evaluates the suitability of these methods for specific dimensions and domains. This framework is applied to generate synthetic datasets that induce changes across the three dimensions of a recent multidimensional LSC framework (SIB; Baes et al., 2024), using examples from psychology, to test and compare the suitability of different methods in detecting these synthetic changes.

Our findings support the hypothesis that recently proposed methods (Valence index, Arousal index, Breadth score; Baes et al., 2024) detect synthetic changes on the SIB dimensions. Control analyses, which adhered to computational linguistics standards (Dubossarsky et al., 2017, 2019), confirmed the absence of these effects in shuffled controls. The implications of these findings are twofold. The ability of SIB methods to detect changes when introducing silver-label synthetic data validates their sensitivity and reliability in detecting and measuring variations in SIB, even in controlled, artificial environments. This, in turn, validates the synthetic LLM-generated sentences in our ICL evaluation suites.

We demonstrated how a synthetic change detection task can assess the sensitivity of various computational approaches, guiding the selection of the most suitable model for specific dimensions and domains. Baes et al.'s (2024) tools, which validated the synthetic SIB datasets, were supported by alternative methods that consistently detected synthetic changes in SIB across all conditions and targets, providing further validation. The Valence index and Sentiment score (ABSA) identified variations in synthetic Sentiment, while Breadth scores (XLL and MPNet) detected increases in synthetic Breadth. Results suggest that these NLP-based methods are more sensitive in detecting synthetic changes than Warriner-based methods, which rely on Valence and Arousal ratings. Future empirical studies on Sentiment and Breadth may consider adopting these NLP models, either as replacements for, or in addition to, existing methods.

Notably, when computing the general LSC score using the SOTA LSC model XLL (Cassotti et al., 2023), it was not sensitive to detecting Sentiment and Intensity. Although XLL shows some sensitivity to identifying synthetic increases in Breadth, it registers a more substantial change when the Breadth score is adjusted according to the method introduced by Baes et al. (2024). It uses the within-bin average cosine distance of target containing sentences as a proxy for the expansion (broadening) or contraction (narrowing) of a word's contextual usage. The inability of XLL to detect the affective dimensions of LSC highlights the necessity of evaluating SOTA models before deploying them in new domains. Future research should investigate whether this weakness in detecting affective dimensions is specific to XLL or extends to other contextualized models in more corpora. This inquiry is particularly salient given recent advances in analyzing fine-grained, continuous semantic shifts through "diachronic word similarity matrices using fast and lightweight word embeddings over arbitrary time periods" (Kiyama et al., 2025).

Findings highlight the need to include affective and connotational aspects of meaning in studies of LSC. In particular, future studies must consider emotional meaning in language models. While psychology has extensively used language to analyze emotion semantics (Jackson et al., 2022; Boyd and Schwartz, 2021), advances in NLP are still exploring how to build models that incorporate sentiment (Goworek and Dubossarsky, 2024) and detect emotion (Mohammad, 2021). Further research is required to detect affective states from text given the cultural and universal aspects of emotion semantics (Jackson et al., 2019). These findings also have implications for existing multidimensional frameworks of LSC (Baes et al., 2024; de Sá et al., 2024) as the evaluation framework provides experimental settings in which to compare the sensitivity of methods to detecting synthetic changes on specific dimensions and domains in a variety of disciplines.

6 Conclusion

The current study introduced a novel general domain evaluation framework. Its three-stage pipeline involves: 1) developing a scalable methodology for generating LLM-based synthetic datasets with silver labels that simulate changes in kinds of LSC; 2) using these datasets to evaluate the relative sensitivity of computational approaches in a synthetic change detection task; and 3) identifying the most suitable method for detecting synthetically induced changes across specific dimensions and domains. We applied this framework to a set of psychological terms. Findings not only supported the validity of proposed computational methods for measuring changes in SIB, but also established a controlled experimental standard for rigorously evaluating existing LSC detection methods and exploring alternative computational approaches. This work is crucial for addressing the substantial gap created by the lack of historical benchmark datasets, which has previously hindered the standardization of metrics and fair comparison of methods. While this innovation benefits all disciplines (e.g., biomedicine, law, theology), it is particularly valuable in the social sciences and humanities, where unique methods are often required to measure complex constructs.

Limitations

Limitations inform future directions. Evaluating the quality of LLM prompt and demonstration examples in the few-shot ICL paradigm is challenging. As LLM evaluation standards are developed (Chang et al., 2024; Ziems et al., 2024), future research might explore automated strategies such as updating prompts based on examples (DSPy)[14] or comparing LLM output from different prompts using a free, unified interface.[15] LLM choice in the evaluation pipeline could be expanded to include open-source models (e.g., FlanT5-XL, Mistral-7B, Mixtral-8x7B), enhancing its accessibility.

Furthermore, our study benefited from using GPT-4o, which is trained on US English and is therefore well-suited for analyzing texts within the Western-centric domain of psychology. However, the cultural and linguistic biases of LLMs may pose challenges for adapting our evaluation pipeline to other languages (Havaldar et al., 2023), although few-shot ICL has proven effective in low-resource languages (Cahyawijaya et al., 2024). Despite the tendency of LLM training data to skew towards the recent past, it successfully generated high-quality sentences that spanned a 1970 to 2019 time period. Future research should focus on refining these models to support broader application across various cultural contexts, languages, and historical periods.

The conceptualization of semantic Breadth is complex and contested. Linguistic definitions suggest breadth encompasses subtypes (e.g., specialization as a subtype of narrowing; Campbell, 2013) highlighting its intricate nature. Given this complexity, it is essential to compare the current measure, which is based on mean within-bin variability of target-containing sentences, with other methods assessing breadth through senses, topics, or prototypical changes: modulations based on literal similarity (Geeraerts, 1997). Future research should investigate whether these measures can detect polysemy's emergence or merely prototype-based modulations of existing concepts.

The synthetic breadth dataset used in this study was constructed using a replacement strategy that may include contextually irrelevant donor contexts. To enhance simulation quality, we propose a three-step validation pipeline: First, select validation models based on performance against a gold-standard dataset, as determined by the highest F1-score from 5-fold stratified K-fold cross-validation. Second, use a probability ratio check with a Masked Language Model (e.g., BioBERT, RoBERTa-large, DeBERTa-v3-large) to confirm the plausibility of replacing donors with target terms, approving sentences that meet a specific probability threshold. Third, ensure semantic alignment through cosine similarity validation with models such as MiniLM-L12-v2 or DistilRoBERTa-v1 Sentence-T5, approving sentences that exceed a set threshold. This process aims to expand the target term's semantic scope while maintaining specificity, but may exclude many sentences. Integrating de Sá et al.'s (2024) ICL approach to simulate Breadth—first teaching the model to disambiguate word senses—could offer an efficient alternative.

Furthermore, the present study does not specify which sense of the term is semantically expanded. Attempting to integrate senses into the synthetic data generation pipeline may provide richer insights. While the specialized psychology corpus and target words exhibit limited senses, general domain corpora introduce ambiguous contexts (e.g., economic sense of "depression"). Notably, current methods for word sense disambiguation may not integrate with distributional approaches as historical linguists do not treat LSC as a set of senses.

Although a body of work estimates valence from natural language, less research has examined the Intensity dimension (Hoemann et al., 2025). In the present study, this restricted the external validation of the Arousal index (Baes et al., 2024), highlighting the need for empirical research in this direction. Furthermore, we must examine the conceptual/terminological link between arousal and hyperbole (i.e., a linguistic form describing a rhetorical, discursive phenomenon like irony) to understand arousal's relation to hyperbole (Burgers et al., 2016; Peña and Ruiz de Mendoza, 2017).

Finally, future research should use the evaluation framework to generate synthetic datasets, and to explore methods, for detecting the Relation dimension (metaphor/metonymy) as highlighted by de Sá et al. (2024). Incorporating the qualitative types of metaphor and metonymy into the empirical study of multidimensional LSC could provide a more comprehensive understanding of LSC, particularly for some domains. Examining how Relation relates to SIB may deepen our understanding of LSC processes by exploring how cognitive principles contribute to semantic innovations.

[14] https://dspy.ai

[15] https://github.com/marketplace/models/azure-openai/gpt-4o/playground

Ethical Considerations

We do not identify any foreseeable risks or potential for harmful use of our work. Analyses containing sentences from the psychology corpus use licensed data that are openly accessible for academic purposes, ensuring transparency and accountability. While LLM-generated sentences are synthetic, they are modeled off samples from the psychology corpus.

Acknowledgments

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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, Ekaterina Vylomova, Michael Zyphur, and Nick Haslam. 2023. The semantic inflation of "trauma" in psychology. *Psychology of Language and Communication*, 27(1):23–45.

Andreas Blank. 1999. Why do new meanings occur? a cognitive typology of the motivations for lexical semantic change. In Andreas Blank and Peter Koch, editors, *Historical semantics and cognition*, pages 61–90. Mouton de Gruter.

Leonard Bloomfield. 1933. *Language*. Compton Printing Works Ltd.

Ryan L Boyd and H Andrew Schwartz. 2021. Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, 40(1):21–41.

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

Christian Burgers, Elly A Konijn, and Gerard J Steen. 2016. Figurative framing: Shaping public discourse through metaphor, hyperbole, and irony. *Communication theory*, 26(4):410–430.

Laura Cabello and Uchenna Akujuobi. 2024. It is simple sometimes: A study on improving aspect-based sentiment analysis performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6597–6610, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Lyle Campbell. 2013. *Historical Linguistics: An Introduction*, ned - new edition, 3 edition. Edinburgh University Press.

Pierluigi Cassotti, Stefano De Pascale, and Nina Tahmasebi. 2024a. 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.

Pierluigi Cassotti, Francesco Periti, Stefano De Pascale, Haim Dubossarsky, and Nina Tahmasebi. 2024b. Computational modeling of semantic change. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–8, St. Julian's, Malta. Association for Computational Linguistics.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. 2023. How many demonstrations do you need for in-context learning? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11149–11159, Singapore. Association for Computational Linguistics.

Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2024. Semantic change characterization with llms using rhetorics. *arXiv preprint arXiv:2407.16624*.

Jader Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2024. Survey in characterization of semantic change. *Preprint*, arXiv:2402.19088.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark. Association for Computational Linguistics.

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

Dirk Geeraerts. 2010. *Theories of lexical semantics*. Oxford University Press.

Andrew Gelman and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY.

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.

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.

Nick Haslam. 2016. Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry*, 27(1):1–17.

Nick Haslam, Ekaterina Vylomova, Michael Zyphur, and Yoshihisa Kashima. 2021. The cultural dynamics of concept creep. *American Psychologist*, 76(6):1013.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, and Haim Dubossarsky. 2021. Challenges for computational lexical semantic change.

Katie Hoemann, Yeasle Lee, Èvelyne Dussault, Simon Devylder, Lyle H. Ungar, Dirk Geeraerts, and Batja Gomes de Mesquita. 2025. The construction of emotional meaning in language.

Joshua Conrad Jackson, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray, and Kristen A. Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Joshua Conrad Jackson, Joseph Watts, Johann-Mattis List, Curtis Puryear, Ryan Drabble, and Kristen A. Lindquist. 2022. From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*, 17(3):805–826. PMID: 34606730.

Hajime Kiyama, Taichi Aida, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura, and Daichi Mochihashi. 2025. Analyzing continuous semantic shifts with diachronic word similarity matrices. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1613–1631, Abu Dhabi, UAE. Association for Computational Linguistics.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*.

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.

Raphael Merx, Ekaterina Vylomova, and Kemal Kurniawan. 2024. Generating bilingual example sentences with large language models as lexicography assistants. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, Canberra, Australia. Association for Computational Linguistics.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

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.

Saif M Mohammad. 2021. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 323–379. Elsevier.

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

M Sandra Peña and Francisco José Ruiz de Mendoza. 2017. Construing and constructing hyperbole. *Studies in figurative thought and language*, 56:41.

Francesco Periti and Stefano Montanelli. 2024a. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11).

Francesco Periti and Stefano Montanelli. 2024b. Lexical semantic change through large language models: a survey. *ACM Computing Surveys*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

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.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *CoRR*, abs/1811.06278.

Nina Tahmasebi and Thomas Risse. 2017. Finding individual word sense changes and their delay in appearance. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 741–749, Varna, Bulgaria. INCOMA Ltd.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Ekaterina Vylomova, Sean Murphy, and Nick Haslam. 2019. Evaluation of semantic change of harm-related concepts in psychology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 29–34.

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.

Yu Xiao, Naomi Baes, Ekaterina Vylomova, and Nick Haslam. 2023. Have the concepts of 'anxiety'and 'depression'been normalized or pathologized? a corpus study of historical semantic change. *PloS one*, 18(6):e0288027.

Heng Yang, Biqing Zeng, Mayi Xu, and Tianxing Wang. 2021. Back to reality: Leveraging pattern-driven modeling to enable affordable sentiment dependency learning. *CoRR*, abs/2110.08604.

Heng Yang, Chen Zhang, and Ke Li. 2023. Pyabsa: A modularized framework for reproducible aspect-based sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 5117–5122. ACM.

Yuxiang Zhou, Jiazheng Li, Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. 2023. The mystery and fascination of llms: A comprehensive survey on the interpretation and analysis of emergent abilities. *arXiv preprint arXiv:2311.00237*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Corpus Counts of Target Terms

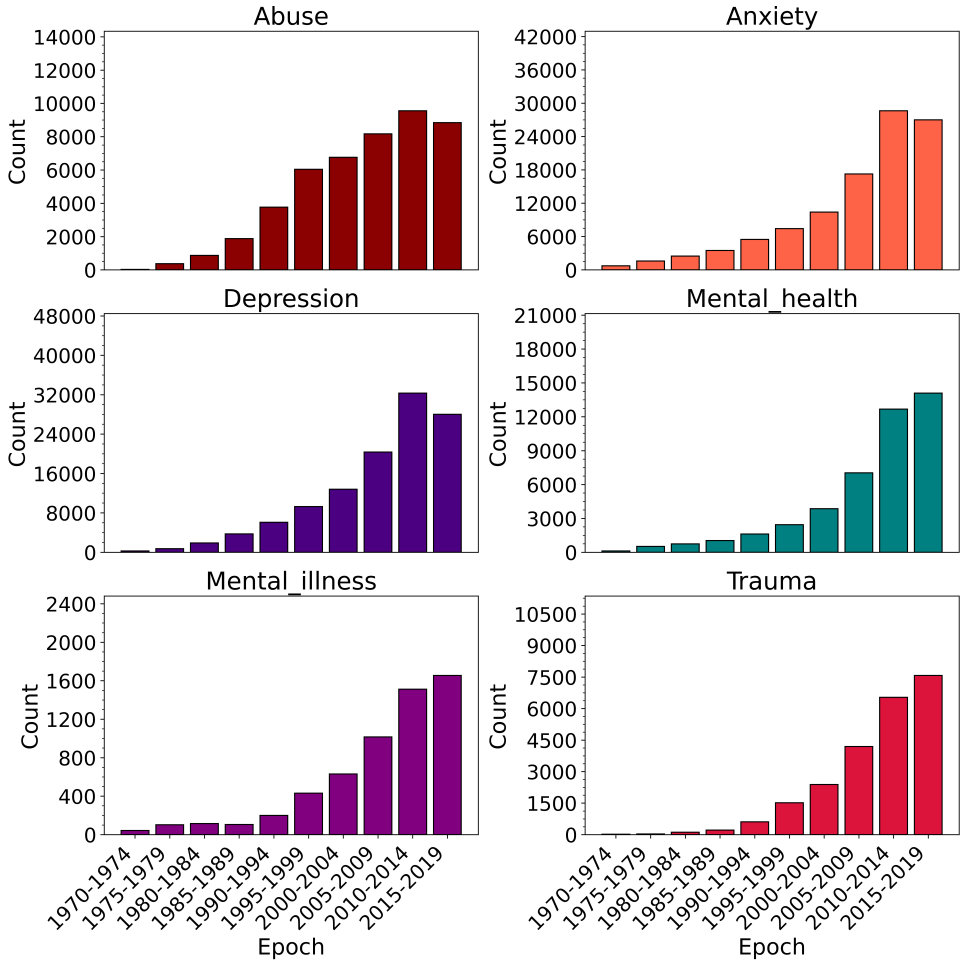

Figure 4: Annual Counts of Sentences where Target Terms Appear in the Psychology Corpus (1970-2019).

B Synthetic Dimension Datasets: Details

Dimension	Target	Neutral (*M*)	Increase (*M*)	Decrease (*M*)	US$
Sentiment	Abuse	5,645 (28)	5,645 (30)	5,645 (29)	17
	Anxiety	9,215 (27)	9,213 (28)	9,213 (28)	28
	Depression	8,828 (27)	8,826 (28)	8,826 (28)	29
	Mental Health	6,348 (28)	6,348 (29)	6,348 (29)	21
	Mental Illness	2,552 (28)	2,552 (28)	2,552 (29)	9
	Trauma	3,563 (28)	3,563 (30)	3,563 (30)	11
Intensity	Abuse	6,802 (28)	6,801 (30)	6,801 (29)	21
	Anxiety	9,659 (26)	9,657 (29)	9,657 (28)	32
	Depression	10,022 (27)	10,020 (30)	10,020 (29)	35
	Mental Health	6,904 (28)	6,899 (32)	6,899 (29)	24
	Mental Illness	2,497 (28)	2,496 (32)	2,496 (29)	10
	Trauma	4,012 (28)	4,012 (30)	4,012 (30)	14
Breadth	Abuse	NA	5,221 (27)	NA	0
	Anxiety	NA	13,635 (26)	NA	0
	Depression	NA	14,463 (27)	NA	0
	Mental Health	NA	14,638 (26)	NA	0
	Mental Illness	NA	14,639 (26)	NA	0
	Trauma	NA	14,650 (26)	NA	0

Table 3: Descriptives for Synthetic Dimension Datasets: Sentence Counts, Sentence Lengths, and Total Generation Cost.

M = Mean Sentence Length of Dataset. Neutral = Neutral, unaltered, input sentences. Increase = Increase on the Dimension of interest. Decrease = Decrease on the Dimension of interest.

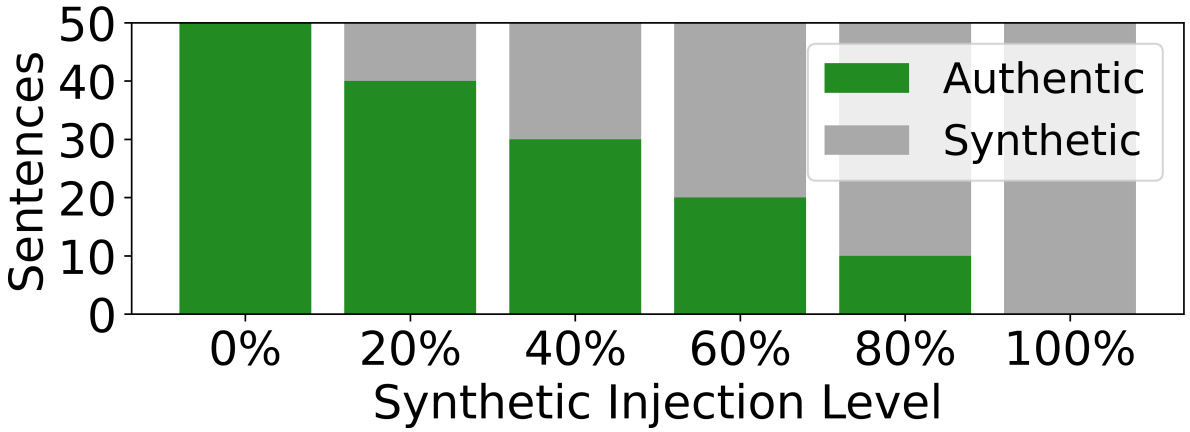

Figure 5: Distribution of Sentences in Each Sample of 50 Sentences.

Dimension	Target	Neutral	Increased Variation	Decreased Variation
Sentiment	Abuse	Child **abuse** is not a single faceted phenomenon.	Child **abuse** is a deeply complex phenomenon that can spur important dialogues and reforms.	Child **abuse** is a multifaceted atrocity with far-reaching and damaging consequences.
	Anxiety	Typical worship reinforces pathologies of **anxiety** and self-deception.	Typical worship empowers resilience in the face of **anxiety** and self-deception.	Typical worship deepens the pathologies of **anxiety** and self-deception.
	Depression	The expression masked **depression** is not a lucky one.	The expression masked **depression** may offer an insightful perspective.	The expression masked **depression** is unfortunately an unsettling one.
	Mental Health	Two views of holiness and its bearing on **mental_health** are discussed.	Two perspectives on holiness and its supportive impact on **mental_health** are discussed.	Two views of holiness and its potential pressure on **mental_health** are discussed.
	Mental Illness	The results suggest that physical or **mental_illness** may decrease creativity.	The results suggest that overcoming physical or **mental_illness** may lead to increased creativity.	The results suggest that physical or **mental_illness** may significantly hinder creativity.
	Trauma	Psychic **trauma** interferes with the normal structuring of experience.	Psychic **trauma** challenges individuals in a way that can lead to the reorganization and enrichment of their experience.	Psychic **trauma** disrupts and fragments the normal structuring of experience.
Intensity	Abuse	Theorists and practitioners alike believe that emotional **abuse** exists.	Theorists and practitioners alike fervently believe that pervasive emotional **abuse** exists.	Theorists and practitioners alike casually believe that subtle emotional **abuse** exists.
	Anxiety	Teacher reported **anxiety** was related to worse time production.	Teacher reported severe **anxiety** was related to significantly worse time production.	Teacher reported mild **anxiety** was related to slightly worse time production.
	Depression	Maternal **depression** continues to play a role in children's development beyond infancy.	Severe maternal **depression** continues to play a profound role in children's development beyond infancy.	Mild maternal **depression** continues to play a subtle role in children's development beyond infancy.
	Mental Health	Eveningness is related to negative physical and **mental_health** outcomes.	Eveningness is alarmingly related to severe negative physical and troubling **mental_health** outcomes.	Eveningness is mildly related to some negative physical and **mental_health** outcomes.

Continued on next page

Dimension	Target	Neutral	Increased Variation	Decreased Variation
	Mental Illness	Biblical and theological considerations underline the importance of the problem about **mental_illness**, but do not provide a solution.	Biblical and theological considerations underline the immense importance and complexity of the problem about **mental_illness**, but do not provide a definitive solution.	Biblical and theological considerations highlight the importance of the issue regarding **mental_illness**, but do not provide a clear solution.
	Trauma	Childhood **trauma** is a key risk factor for psychopathology.	Childhood **trauma** is a critical and devastating risk factor for severe psychopathology.	Childhood **trauma** is a notable but moderate risk factor for mild psychopathology.
Breadth	Abuse	Sexual **exploitation** is an expression of a power relationship.	Sexual **abuse** is an expression of a power relationship.	NA
	Anxiety	Adolescents' **state of mind** with regard to attachment and representations regarding separation were examined.	Adolescents' **anxiety** with regard to attachment and representations regarding separation were examined.	NA
	Depression	Iranian college students showed more **anxiety** than their British peers.	Iranian college students showed more **depression** than their British peers.	NA
	Mental Health	Such a scale may alert clinicians early in treatment to issues related to **trauma**	Such a scale may alert clinicians early in treatment to issues related to **mental_health**	NA
	Mental Illness	Excessive estrogen influence produces anxiety, **agitation**, irritability, and lability.	Excessive estrogen influence produces anxiety, **mental_illness**, irritability, and lability.	NA
	Trauma	Further investigation of pathological **dissociation** in Hong Kong is necessary.	Further investigation of pathological **trauma** in Hong Kong is necessary.	NA

Table 5: Sample of Short Synthetic Sentences from the Synthetic Datasets for each Target term.

C In-Context Learning Paradigm

The study generated synthetic datasets to simulate changes in Sentiment and Intensity using 36,151 and 39,896 neutral baseline sentences, respectively. Neutral sentences were sampled by linking words in each sentence with their mean valence or arousal scores from the NRC-VAD lexicon (0-1) (Mohammad, 2018) and filtering by a dynamic range. This neutral range is adjusted from the median of each dataset by ±0.01, targeting 25th-75th percentile bounds or 500-1500 unique sentences per epoch. See Figures 6 and 7 for a breakdown of neutral sentence counts per epoch provided as input to the LLM using the prompts below.

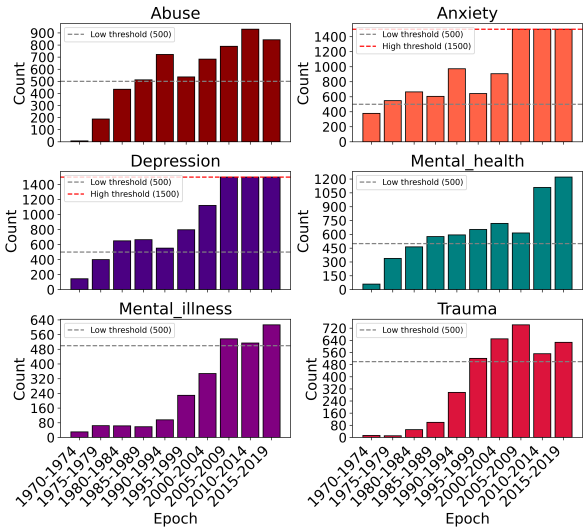

Figure 6: Counts of Neutral Sentences (valence Scores).

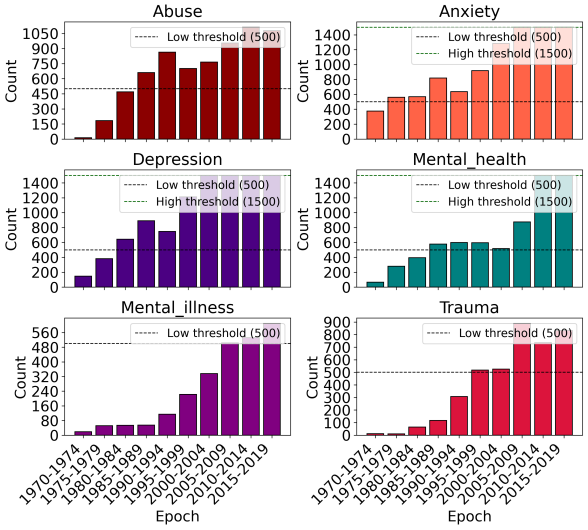

Figure 7: Counts of Neutral Sentences (Arousal Scores).

For each neutral sentence, one inference call to GPT-4o is made through the OpenAI API to generate variations of increased and decreased Sentment or Intensity. Only the samples for *anxiety* and *depression* reached the upper limit of 1,500 sentences for the final three epochs, while other targets did not exceed 500 sentences per epoch (allowing for unique sentences across each of the 10 iterations of up to 50 unique sentences). The sentence generation prioritized quality and maintained a neutral baseline to allow for adequate variation.

The ChatGPT API with a temperature setting of 1.00 was used to ensure semantic accuracy and prevent errors (Periti and Montanelli, 2024b), while allowing for a balance between deterministic and creative responses. Note that there were challenges in maintaining target terms in the sentences, particularly for positive sentiment variations. Fewer manual adjustments were needed for Intensity than Sentiment. GPT-4o struggled to vary 97% of the sentences to contain more positive sentiment for *abuse* (28), *anxiety* (110), *depression* (46), *mental_health* (1), *trauma* (2) as it replaced targets with positive terminology against instructions. For Intensity data, fewer sentences required manual alteration: only for *abuse* (4), *depression* (2), *mental_health* (2), *trauma* (1). Rows (196) were detected and manually altered to retain the target term while ensuring variation in the dimension relative to the neutral sentence. The final validated datasets, detailed in Table 6, are available on the GitHub repository: [MASKED LINK].

Target	Dimension	Neutral	Increase	Decrease	US$
abuse	Sentiment	5,645	5,645	5,645	17
	Intensity	6,802	6,801	6,801	21
anxiety	Sentiment	9,215	9,213	9,213	28
	Intensity	9,659	9,657	9,657	32
depression	Sentiment	8,828	8,826	8,826	29
	Intensity	10,022	10,020	10,020	35
mental_health	Sentiment	6,348	6,348	6,348	21
	Intensity	6,904	6,899	6,899	24
mental_illness	Sentiment	2,552	2,552	2,552	9
	Intensity	2,497	2,496	2,496	10
trauma	Sentiment	3,563	3,563	3,563	11
	Intensity	4,012	4,012	4,012	14

Table 6: Sentence counts and Cost for Synthetic Sentiment and Intensity Datasets.

Prompt for Synthetic *Sentiment*

PROMPT_INTRO = """ In psychology research, 'Sentiment' is defined as "a term's acquisition of a more positive or negative connotation." This task focuses on the sentiment of the term **«target_word»**.

Task
You will be given a sentence containing the term **«target_word»**. Your goal is to write two new sentences:
1. One where **«target_word»** has a **more positive connotation** (enclose this sentence between '<positive target_word>' and '</positive target_word>' tags).
2. One where **«target_word»** has a **more negative connotation** (enclose this sentence between '<negative target_word>' and '<negative target_word>' tags).

Rules
1. The term **«target_word»** must remain **exactly as it appears** in the original sentence:
- Do **not** replace, rephrase, omit, or modify it in any way.
- Synonyms, variations, or altered spellings are not allowed.
2. **Meaning and Structure**:
- Stay true to the original context and subject matter.
- Maintain the sentence's structure and ensure grammatical accuracy.
3. **Sentiment Adjustments**:
- **Positive Sentiment**: Reflect strengths or benefits realistically, while respecting the potential negativity of **«target_word»**.
- **Negative Sentiment**: Highlight risks or harms appropriately, avoiding exaggeration or trivialization.

Important
- Any response omitting, replacing, or altering **«target_word»** will be rejected.
- Ensure the output is:
- **Grammatically correct**
- **Sensitive and serious** in tone
- **Free from exaggeration or sensationalism**
- **Strictly following the XML-like tag format for sentiment variations**

Follow these guidelines strictly to produce valid responses. """

Prompt for Synthetic *Intensity*

PROMPT_INTRO = """In psychology research, Intensity is defined as "the degree to which a word has emotionally charged (i.e., strong, potent, high-arousal) connotations." This task focuses on the intensity of the term **«target_word»**.

Task
You will be given a sentence containing the term **«target_word»**. Your goal is to write two new sentences:
1. One where **«target_word»** is **less intense** (enclose this sentence between '<decreased target_word intensity>' and '</decreased target_word intensity>' tags).
2. One where **«target_word»** is **more intense** (enclose this sentence between '<increased target_word intensity>' and '</increased target_word intensity>' tags).

Rules
1. The term **«target_word»** must remain **exactly as it appears** in the original sentence:
- Do **not** replace, rephrase, omit, or modify it in any way.
- Synonyms, variations, or altered spellings are not allowed.
2. **Meaning and Structure**:
- Stay true to the original context and subject matter.
- Maintain the sentence's structure and ensure grammatical accuracy.

Important
- Any response omitting, replacing, or altering **«target_word»** will be rejected.
- Ensure the output is:
- **Grammatically correct**
- **Sensitive and serious** in tone
- **Free from exaggeration or sensationalism**
- **Strictly following the XML-like tag format for intensity variations**

Follow these guidelines strictly to produce valid responses. """

D Demonstration Examples: Synthetic Sentiment

Target	Neutral	Positive Sentiment	Negative Sentiment
Abuse	Child **abuse** is most likely to occur when socially isolated parents react impulsively to aversive stimuli emitted by their children.	Child **abuse** is less likely to occur when socially isolated parents respond lovingly to their children's behavior.	Child **abuse** is most likely to occur when socially isolated parents react aggressively to their children's challenging behavior.
Abuse	The children represented a wide spectrum of sexual **abuse**.	The children represented a meaningful spectrum of sexual **abuse**.	The children represented a devastating spectrum of sexual **abuse**.
Abuse	Euphoric properties of cocaine lead to the development of chronic **abuse**, and appear to involve the acute activation of central DA neuronal systems.	Euphoric properties of cocaine lead to the growth of chronic **abuse**, and appear to involve the acute activation of central DA pleasure systems.	Emotional properties of cocaine lead to the decline into chronic **abuse**, and appear to involve the acute activation of central DA pain systems.
Abuse	Substance abuse helps the individual deal with distress associated with family interactions.	Substance **abuse** helps the individual temporarily cope positively with family interactions.	Substance **abuse** makes the individual endure the overwhelming pain and alienation associated with family interactions.
Abuse	The study determined that 84 of the sample reported a history of **abuse** or neglect.	The study determined that 84 of the sample acknowledged a transformative history of overcoming **abuse** or neglect.	The study determined that 84 of the sample complained of a miserable history of **abuse** or neglect.
Anxiety	Previous work suggests that social **anxiety** is inconsistently related to alcohol use.	Previous work agrees that social **anxiety** is sometimes related to alcohol use.	Previous work warns that social anxiety is unpredictably related to alcohol use.
Anxiety	A small yet emerging body of research on the relationship between **anxiety** and driving suggests that higher levels of state anxiety may lead to more dangerous driving behaviors.	A small yet emerging body of research on the positive relationship between **anxiety** and driving suggests that higher levels of state anxiety may lead to more daring driving behaviors.	A small yet emerging body of research on the problematic relationship between **anxiety** and driving suggests that more disturbing levels of state **anxiety** may lead to more disastrous driving behaviors.
Anxiety	Findings suggest that individuals high in **anxiety** show greater contextual fear generalization as measured by US expectancy.	Findings suggest that individuals high in **anxiety** show greater contextual concern generalization as measured by US hope.	Findings suggest that individuals high in **anxiety** show greater contextual terror generalization as measured by US dread.
Anxiety	General **anxiety** and evoked imagery of death as a person were measured in 75 male Catholic college students and seminarians.	General **anxiety** and vivid imagery of hope as a person were measured in 75 male Catholic college students and seminarians.	General **anxiety** and frightening imagery of death as a person were measured in 75 male Catholic college students and seminarians.

Continued on next page

Target	Neutral	Positive Sentiment	Negative Sentiment
Anxiety	Results indicated that emotion dysregulation significantly mediated the relationship between child abuse severity and attachment-related **anxiety** and avoidance.	Results indicated that emotion variation positively mediated the relationship between childhood experiences and attachment-related **anxiety** and care.	Results indicated that emotion disturbance problematically mediated the relationship between child abuse severity and attachment-related **anxiety** and terror.
Depression	The present study was conducted to test predictions derived from the hypothesis that **depression** may serve the purpose of adaptively facilitating disengagement from obsolete cognitive plans.	The present study was conducted to test predictions derived from the hypothesis that **depression** may serve the purpose of helping people make better cognitive plans.	The present study was conducted to test predictions derived from the hypothesis that **depression** may prevent people from carrying out destructive cognitive plans.
Depression	Vision loss was a consistent predictor of both onset and persistence of **depression**, even after a wide range of covariates had been adjusted.	Vision loss was a positive predictor of both beginning and retaining **depression**, even after a wide range of covariates had been included.	Vision loss was an unavoidable predictor of both suffering and enduring **depression**, even after a wide range of covariates had been controlled.
Depression	This study examined whether distinct groups of young adolescents with mainly anxiety or mainly **depression** could be identified in a general population sample.	This study examined whether unique groups of young adolescents with mainly vigilance or mainly **depression** could be identified in a general population sample.	This study examined whether pathological groups of young adolescents with mainly fear or mainly **depression** could be isolated in a general population sample.
Depression	In most people with recurrent **depression**, mindfulness skills are expressed evenly across different domains.	In most people who live with **depression**, mindfulness skills are expressed in a balanced way across different domains.	In most people who struggle with untreatable **depression**, mindfulness habits are expressed monotonously across different domains.
Depression	The aim of the study was to test the effect of differing information regarding the rationale given to participants for a study on **depression** symptoms.	The hope of the study was to test the effect of diverse information regarding the clarifying reasons bestowed on participants for an exploration of **depression** features.	The aim of the study was to test the effect of differing information regarding the dreary explanation given to participants for a study on **depression** pathologies.
Mental Health	This paper maintains that **mental_health** delivery systems must be supplemented by critical analyses of the hidden assumptions that guide policy and technique decisions.	This paper hopes that **mental_health** delivery systems must be improved by enlightened analyses of the hidden assumptions that lead beneficial policy and technique decisions.	This paper warns that **mental_health** delivery systems must be supplemented by harsh analyses of the deep-seated errors that undermine policy and technique decisions.

Continued on next page

Target	Neutral	Positive Sentiment	Negative Sentiment
Mental Health	The federal regulations governing confidentiality of alcohol and drug abuse patient records are examined with respect to their applicability to **mental_health** and other medical records.	The federal regulations protecting confidentiality of alcohol and drug use records are examined with respect to their applicability to **mental_health** and other well-being records.	The federal regulations restricting access to alcohol and drug abuse patient records are examined with respect to their potential shortcomings for **mental_health** and other medical records.
Mental Health	Young people are particularly vulnerable to unemployment and the consequences of this for psychosocial development and **mental_health** are not well understood.	Young people are particularly responsive to leisure and the consequences of this for psychosocial well-being and **mental_health** will benefit from more understanding.	Young people are particularly vulnerable to unemployment and the threats of this for dysfunction and **mental_health** are poorly understood.
Mental Health	This study suggests that the long-term outcome in schizophrenic patients followed by a community-based **mental_health** service is generally poor and multifaceted.	This study suggests that the long-term improvement in people with schizophrenia followed by a community-based **mental_health** service is generally variable.	This study warns that the long-term outcome in schizophrenic patients followed by a community-based **mental_health** clinical is generally poor and incoherent.
Mental Health	The stigma of having psychological problems is a barrier to seeking **mental_health** treatment, but little research has examined whether this stigma influences the experiences of those in treatment.	The public image of having well-being challenges is a bridge to seeking **mental_health** help, but little research has examined whether this image influences the experiences of those in care.	The shame of having psychological illness is an obstacle to seeking **mental_health** treatment, but little research has examined whether this shame increases the misery of those in treatment.
Mental Illness	Internet addiction (IA) is an emerging social and **mental_health** issue among youths.	Internet engagement (IE) is a rising social and **mental_health** issue among youths.	Internet addiction (IA) is a looming social and **mental_health** disorder among youths.
Mental Illness	Second, we asked to what extent suicides of older mentally ill persons are definitely created by their **mental_illness**.	Second, we asked to what extent suicides of older persons are definitely created by their **mental_illness**.	Second, we asked to what extent suicides of older mentally ill persons are definitely made worse by their **mental_illness**.
Mental Illness	It was found that rejection of the mentally ill in situations of social relations was linked to prior personal experience with **mental_illness**, perceived dangerousness of the mentally ill, and age of the survey respondent.	It was found that welcoming of people in situations of social relations was linked to prior positive personal experience with **mental_illness**, perceived safety of these people, and age of the survey respondent.	It was found that rejection of the mentally ill in situations of social relations was linked to negative prior personal experience with **mental_illness**, perceived dangerousness of the mentally ill, and age of the survey respondent.
Mental Illness	In over 50 of cases continuation of in-patient stay was necessitated by the severity of **mental_illness**.	In over 50 of cases continuation of stay in care was necessitated by the level of **mental_illness**.	In over 50 of cases being restricted to hospital was necessitated by the severity of **mental_illness**.

Continued on next page

Target	Neutral	Positive Sentiment	Negative Sentiment
Mental Illness	Much controversy exists over the treatment of **mental_illness** and many critics argue that the exercise of medical authority results in the social control of the mentally ill.	Much conversation exists over the care of **mental_illness** and many writers argue that the medical authorities enhance the social enhancement of mental health.	Much disagreement exists over the treatment of **mental_illness** and many critics argue that the abuse of medical tyranny results in the domination of the mentally ill.
Trauma	This paper presents a cognitive-behavioral model for conceptualizing and intervening in the area of sexual **trauma**.	This paper celebrates a cognitive-behavioral model for promoting new ideas and helping in the area of sexual **trauma**.	This paper presents a cognitive-behavioral model for thinking about and wresting with the harmful problem of sexual **trauma**.
Trauma	In most classrooms in most schools, there are students who have suffered complex **trauma** who would benefit from a system-wide, trauma-informed approach to schooling.	In most classrooms in most schools, there are students who have experienced complex **trauma** who would benefit from a system-wide, responsive and enlightened approach to schooling.	In most classrooms in most schools, there are students who have suffered damaging **trauma** whose problems need a system-wide, illness-based approach to schooling.
Trauma	Research has shown that women are more likely to develop PTSD subsequent to **trauma** exposure in comparison with men.	Research has shown that women are more likely to develop PTSD subsequent to **trauma** experiences in comparison with men.	Research has shown that women are more likely to deteriorate into PTSD subsequent to **trauma** exposure in comparison with men.
Trauma	Numerous homeless youth experience **trauma** prior to leaving home and while on the street.	Numerous resilient youth learn to navigate **trauma** prior to leaving home and while adapting to life on the street.	Numerous homeless youth endure significant **trauma** prior to leaving home and while facing severe challenges on the street.
Trauma	The meaning of **trauma** within psychology has for a long time been viewed mostly from a pathologizing standpoint.	The meaning of **trauma** within psychology has for a long time needed to be viewed from a more compassionate and strengths-based standpoint.	The meaning of **trauma** within psychology has for a long time been viewed mostly from a negative and overly disease-focused standpoint.

Table 8: Expert Crafted Sentiment Variations for Neutral Sentences for inference calls to GPT-4o for the Few-Shot ICL Paradigm.

E Demonstration Examples: Synthetic Intensity

Target	Neutral	High Intensity	Low Intensity
Abuse	Clinically, however, individual questions that use broad labeling terms are more likely to identify women as having a history of **abuse**.	Clinically, however, individual questions that use extreme labeling terms are more likely to reveal women as having a severe history of **abuse**.	Clinically, however, individual questions that use broad labeling terms are more likely to identify women as having a mild history of **abuse**.
Abuse	Most care workers said that they would be willing to report **abuse** anonymously.	Most care workers cried that they would be delighted to report extreme instances of **abuse** anonymously.	Most care workers said that they would be willing to report trivial **abuse** anonymously.
Abuse	There is greater emphasis on recognizing that older people may be subjected to **abuse** and neglect by family members and the community as well.	There is a significant emphasis on recognizing that older people may be subjected to severe **abuse** and appalling neglect by family members and the community as well.	There is some emphasis on recognizing that older people may experience weak **abuse** by family members and the community as well.
Abuse	Education on financial **abuse** for both elders and their adult children and establishment of income support programs are urgently needed.	Education on ordinary financial **abuse** for both elders and their adult children and urgent establishment of income support programs are desperately needed.	Education on financial **abuse** for both elders and their adult children and establishment of income support programs will occur.
Abuse	There was no association between physical **abuse** and depressive symptoms through either self-compassion or gratitude.	There was no association between frightening physical **abuse** and cold symptoms through either emotional contagion or extreme gratitude.	There was no association between mild physical **abuse** and state of mind through either complacency or gratitude.
Anxiety	The spread of **anxiety** as seen in curves of generalization seems greater at the unconscious than at the conscious level.	The uncontrollable spread of intense **anxiety** as seen in spikes of generalization seems more vivid at the unconscious than at the conscious level.	The spread of mild **anxiety** as seen in curves of generalization seems greater at the unconscious than at the conscious level.
Anxiety	These findings suggest that two important factors to be considered by researchers, educators, and mental_health professionals are adults' perceptions of their fathers' level of acceptance-rejection and the amount of **anxiety** they experience in their relationship with God.	These findings cry out that two powerful factors to be considered by researchers, educators, and mental_health professionals are adults' perceptions of their fathers' extreme level of rejection and the intense amount of **anxiety** they experience in their relationship with God.	These findings suggest that two important factors to be considered by researchers, educators, and other professionals are adults' perceptions of their fathers' level of acceptance and the amount of mild **anxiety** they experience in their relationship with God.

Continued on next page

Target	Neutral	High Intensity	Low Intensity
Anxiety	Self-compassion might be an alternative strategy for cognitive reappraisal in the management of shame-proneness and social **anxiety**.	Emotion exaggeration might be an alternative strategy for overcoming upset in the management of shame and extreme social **anxiety**.	Meditation might be an alternative strategy for cognitive reappraisal in the management of boredom and mild social **anxiety**.
Anxiety	The chronic **anxiety** level of the subject may be related to the ease of acquisition and spread of new anxiety responses.	The intense **anxiety** level of the subject may be related to the ease of acquisition and catastrophic spread of extreme anxiety responses.	The mild **anxiety** level of the subject may be related to the ease of acquisition and generalization of new responses.
Anxiety	Results indicated that greater attachment **anxiety** and avoidance were linked to lower levels of life satisfaction in both gay men and lesbians.	Results cried out that extreme attachment **anxiety** and avoidance were linked to desperate levels of life misery in both gay men and lesbians.	Results indicated that attachment **anxiety** and peacefulness were linked to lower levels of life satisfaction in both gay men and lesbians.
Depression	A combined medical and psychiatric treatment of a **depression** consequent to a colostomy and an organic impotence following rectal resection for cancer in a 33-year-old man has been described.	A combined medical and psychiatric treatment of an intense **depression** consequent to a colostomy and a severe organic impotence following surgical rectal tissue destruction for cancer in a 33-year-old man has been described.	A combined medical and psychiatric treatment of a mild **depression** consequent to a colostomy and an organic impotence following rectal resection for cancer in a 33-year-old man has been described.
Depression	A 35-year-old woman had a history of increasing irritability and liability to attacks of **depression** related to a complete inability to have coital orgasms.	A 35-year-old woman had a fearsome history of crescendoing irritability and liability to severe attacks of **depression** related to a horrendous inability to have coital orgasms.	A 35-year-old woman had a history of sleepiness and liability to periods of mild **depression** related to an inability to have coital orgasms.
Depression	During acute asthma these appear to be radically altered into sadness and longing, and subjected to generalized inhibition similar to that seen in states of **depression**.	During severe, life-threatening asthma episodes these appear to be radically altered into intense misery, and subjected to generalized inhibition similar to that seen in states of extreme **depression**.	During asthma these appear to be altered into boredom and tiredness, and subjected to generalized inhibition similar to that seen in states of low-level **depression**.
Depression	Differences in response in the same individual seem related to mood and attitude as well as to transient stress, with the response being lower on days of **depression**.	Scary differences in response in the same individual seem related to intense mood and attitude as well as to sudden stress, with the emotional response being more intense on days of destructive **depression**.	Predictable differences in response in the same individual seem related to mood, attitude and life experiences, with the subdued response being mild on days of everyday **depression**.

Continued on next page

Target	Neutral	High Intensity	Low Intensity
Depression	The **depression** was treated by the introduction of behaviors incompatible with the **depression**.	The intense depression was treated by the shocking introduction of uncontrollable behaviors incompatible with the severe **depression**.	The mild **depression** was treated by the introduction of behaviors incompatible with it.
Mental Health	Community **mental_health** espouses an innovative conception for psychological services in the university community.	Community **mental_health** fights for a divisive conception for psychological services in the overwhelmed university community.	Community **mental_health** espouses a dull conception for services in the university community.
Mental Health	We also opine that if restraints are misused by **mental_health** or child welfare treatment settings, then their misuse may be considered a subject of a patient maltreatment, abuse, criminal or civil action.	We also exclaim that if harsh restraints are abused by **mental_health** or child welfare treatment settings, then their damaging misuse may be criticized as a subject of extreme patient maltreatment, abuse, criminal or civil action.	We also state that if restraints are used by **mental_health** or child welfare treatment settings, then they may be considered a subject of a discussion.
Mental Health	This research is a secondary data analysis of the impact of adolescents' mental/substance-use disorders and dual diagnosis on their utilization of drug treatment and **mental_health** services.	This research is an intense data analysis of the terrible impact of adolescents' mental/substance abuse disorders and severe compounding problems on their abuse of drug treatment and **mental_health** services.	This research is a data analysis of the impact of adolescents' experiences on their utilization of normal treatment and **mental_health** services.
Mental Health	The findings emphasize the need for family-based treatment for CP that addresses parent behaviors and adolescent **mental_health**.	The findings make a heartfelt plea for the desperate need for family-based treatment for CP that challenges destructive parent behaviors and adolescent **mental_health** diseases.	The findings summarize the need for family-based treatment for CP that addresses ordinary parent behaviors and mild adolescent **mental_health**.
Mental Health	Our findings suggest that maternal **mental_health** influences child sleep behavior at 18 months after birth, and not vice versa.	Our exciting findings suggest that damaged maternal **mental_health** destructively influences child sleep behavior at 18 months after birth, and not vice versa.	Our findings suggest that ordinary maternal **mental_health** influences child normal sleep behavior at 18 months after birth, and not vice versa.
Mental Illness	Problems of definition and classification in psychiatry and the impact of **mental_illness** on the individual and the community pose unique problems for psychiatric register studies.	Horrible problems of definition and classification in psychiatry and the harsh impact of severe **mental_illness** on the individual and the community pose frightening problems for psychiatric register studies.	Issues of definition and classification in psychiatry and the impact of mild **mental_illness** on the individual and the community arise in register studies.

Continued on next page

Target	Neutral	High Intensity	Low Intensity
Mental Illness	In parents and collateral relatives of the autistic children, 3.2% had a serious **mental_illness**, and 4.8% of siblings were markedly abnormal.	In desperate parents and relatives of the severely autistic children, 3.2% had a serious **mental_illness**, and 4.8% of siblings were extremely abnormal.	In parents and relatives of the mildly autistic children, 3.2% had an ordinary **mental_illness**, and 4.8% of siblings were normal.
Mental Illness	Consistent with genetic essentialism, genetic attributions increased the perceived seriousness and persistence of the **mental_illness** and the belief that siblings and children would develop the same problem.	Consistent with the horrors of genetic essentialism, genetic attributions exaggerated the perceived severity and uncontrollability of the severe **mental_illness** and the destructive belief that siblings and children would develop the same extreme problem.	Consistent with genetic essentialism, genetic attributions influenced views about the **mental_illness** and the belief that siblings and children would develop it.
Mental Illness	The target population was urban, homeless, HIV+ individuals with substance dependence and/or **mental_illness** diagnoses.	The completely overwhelmed target population was urban, homeless, HIV+ individuals with severe substance abuse and/or unmanageable **mental_illness** diagnoses.	The target population was urban, ambulatory, healthy individuals with mild **mental_illness** diagnoses.
Mental Illness	Doctors, including general practitioners, experience higher levels of **mental_illness** than the general population.	Doctors, including general practitioners, experience higher levels of **mental_illness** than the general population.	Doctors, including general practitioners, experience higher levels of **mental_illness** than the general population.
Trauma	They tend to be more liberal in their attitudes toward abortion than women in general; however, women who experienced a greater degree of psychic **trauma** tended to be more conservative in their attitudes.	They tend to be more extremely callous in their attitudes toward the horrors of abortion than women in general; however, women who suffered a greater degree of violent psychic **trauma** tended to be more fearful in their attitudes.	They tend to be more accepting in their attitudes toward children than women in general; however, women who experienced mild psychic **trauma** tended to be more conservative in their attitudes.
Trauma	The **trauma** was overwhelming.	The intense **trauma** was completely overwhelming.	The mild **trauma** was unproblematic.
Trauma	The choice of defensive style was found related to at least three factors: an early history of **trauma**, especially separation, parental encouragement of toughness, and essentially a counterphobic family style.	The choice of emotional overreaction was found related to at least three factors: an early history of extreme **trauma**, especially harsh abandonment, parental punishment, and essentially an emotionally destructive family style.	The choice of coping style was found related to at least three factors: an early history of mild **trauma**, especially independence, parental encouragement, and essentially a dull and normal family style.

Continued on next page

Target	Neutral	High Intensity	Low Intensity
Trauma	It is an attempt to bring the **trauma** arising from the external world into the internal world and thus to create an illusion of mastery and control.	It is a desperate attempt to bring the unbearable **trauma** threatening from the external world into the internal world and thus to create a poisonous illusion of mastery and control.	It is an attempt to bring the mild **trauma** arising from the external world into the internal world and thus to create a sense of peace and tranquillity.
Trauma	The international standard for setting ski bindings is based on the measurement of the tibia proximal width because of the propensity of this bone to suffer **trauma** as the ski and skier attempt to go in different directions.	The disgraceful international standard for setting ski bindings is based on the measurement of the tibia proximal width because of the scary propensity of this bone to suffer severe **trauma** as the ski and skier attempt to go in different directions.	The international standard for setting ski bindings is based on the measurement of the tibia proximal width because of the propensity of this bone to experience mild **trauma** as the ski and skier attempt to go in different directions.

Table 10: Expert Crafted Intensity Variations for Neutral Sentences for inference calls to GPT-4o for the Few-Shot ICL Paradigm.

F List of Donor Terms: Synthetic Breadth

Target (Synset)	Sibling (Synset)	Lin Similarity	Cosine Similarity
Abuse (abuse.n.02)	Disparagement (disparagement.n.01)	1.54	0.89
	Contempt (contempt.n.03)	1.49	0.86
	Impudence (impudence.n.01)	1.47	0.84
	Ridicule (ridicule.n.01)	1.34	0.91
	Derision (derision.n.01)	1.24	0.81
	Blasphemy (blasphemy.n.01)	1.07	0.89
Abuse (maltreatment.n.01)	Exploitation (exploitation.n.02)	1.78	0.86
	Disregard (disregard.n.02)	1.67	0.82
	Harassment (harassment.n.02)	1.55	0.84
	Annoyance (annoyance.n.05)	1.37	0.83
Anxiety (anxiety.n.01)	Depression (depression.n.01)	2.09	0.91
	Mental Health (mental_health.n.01)	1.85	0.89
	Trauma (trauma.n.02)	1.70	0.90
	Mental Illness (mental_illness.n.01)	1.60	0.92
	Dissociation (dissociation.n.02)	1.55	0.90
	Hypnosis (hypnosis.n.01)	1.43	0.89
	Delusion (delusion.n.01)	1.42	0.89
	Anhedonia (anhedonia.n.01)	1.33	0.84
	Agitation (agitation.n.01)	1.31	0.91
	Depersonalization (depersonalization.n.02)	1.31	0.90
	Irritation (irritation.n.01)	1.26	0.89
	Morale (morale.n.01)	1.26	0.89
	Nervousness (nervousness.n.02)	1.24	0.84
	Enchantment (enchantment.n.02)	1.24	0.92
	Cognitive State (cognitive_state.n.01)	1.21	0.87
	State of Mind (state_of_mind.n.01)	1.21	0.83
	Elation (elation.n.01)	1.15	0.91
	Fugue (fugue.n.02)	1.06	0.91
	Hallucinosis (hallucinosis.n.01)	1.05	0.92
	Abulia (abulia.n.01)	0.97	0.80
Depression (depression.n.01)	Anxiety (anxiety.n.01)	2.09	0.91
	Mental Health (mental_health.n.01)	1.87	0.89
	Trauma (trauma.n.02)	1.71	0.84
	Mental Illness (mental_illness.n.01)	1.61	0.88
	Dissociation (dissociation.n.02)	1.56	0.89
	Morale (morale.n.01)	1.26	0.91
	Depersonalization (depersonalization.n.02)	1.32	0.92
	Enchantment (enchantment.n.02)	1.25	0.88
	Delusion (delusion.n.01)	1.43	0.90
	Hypnosis (hypnosis.n.01)	1.44	0.83
	Anhedonia (anhedonia.n.01)	1.34	0.84
	Agitation (agitation.n.01)	1.32	0.89
	Nervousness (nervousness.n.02)	1.25	0.84
	Cognitive State (cognitive_state.n.01)	1.22	0.85
	State of Mind (state_of_mind.n.01)	1.22	0.80
	Irritation (irritation.n.01)	1.27	0.85
	Fugue (fugue.n.02)	1.07	0.86

Continued on next page

Target (Synset)	Sibling (Synset)	Lin Similarity	Cosine Similarity
	Hallucinosis (hallucinosis.n.01)	1.05	0.89
	Abulia (abulia.n.01)	0.97	0.76
Depression (depression.n.04)	Forlornness (forlornness.n.01)	1.52	0.88
	Sorrow (sorrow.n.02)	1.36	0.86
	Heaviness (heaviness.n.02)	1.15	0.77
	Misery (misery.n.02)	1.10	0.89
	Melancholy (melancholy.n.01)	1.06	0.87
	Sorrow (sorrow.n.01)	1.13	0.85
	Weepiness (weepiness.n.01)	1.02	0.83
	Downheartedness (downheartedness.n.01)	0.93	0.88
	Dolefulness (dolefulness.n.01)	0.84	0.86
Mental Health (mental_health.n.01)	Depression (depression.n.01)	1.87	0.89
	Anxiety (anxiety.n.01)	1.85	0.89
	Trauma (trauma.n.02)	1.55	0.86
	Mental Illness (mental_illness.n.01)	1.46	0.91
	Dissociation (dissociation.n.02)	1.43	0.90
	Hypnosis (hypnosis.n.01)	1.32	0.86
	Delusion (delusion.n.01)	1.31	0.84
	Anhedonia (anhedonia.n.01)	1.24	0.83
	Agitation (agitation.n.01)	1.22	0.90
	Depersonalization (depersonalization.n.02)	1.22	0.87
	Irritation (irritation.n.01)	1.18	0.88
	Morale (morale.n.01)	1.17	0.92
	Nervousness (nervousness.n.02)	1.16	0.84
	Enchantment (enchantment.n.02)	1.16	0.88
	Cognitive State (cognitive_state.n.01)	1.13	0.90
	State of Mind (state_of_mind.n.01)	1.13	0.85
	Elation (elation.n.01)	1.08	0.90
	Fugue (fugue.n.02)	1.00	0.86
	Hallucinosis (hallucinosis.n.01)	0.99	0.88
	Abulia (abulia.n.01)	0.92	0.79
Mental Illness (mental_illness.n.01)	Depression (depression.n.01)	1.61	0.88
	Anxiety (anxiety.n.01)	1.60	0.92
	Trauma (trauma.n.02)	1.36	0.87
	Dissociation (dissociation.n.02)	1.27	0.90
	Hypnosis (hypnosis.n.01)	1.18	0.86
	Delusion (delusion.n.01)	1.18	0.86
	Anhedonia (anhedonia.n.01)	1.12	0.80
	Agitation (agitation.n.01)	1.11	0.88
	Depersonalization (depersonalization.n.02)	1.10	0.88
	Irritation (irritation.n.01)	1.07	0.87
	Morale (morale.n.01)	1.06	0.87
	Nervousness (nervousness.n.02)	1.05	0.80
	Enchantment (enchantment.n.02)	1.05	0.90
	Cognitive State (cognitive_state.n.01)	1.03	0.86
	State of Mind (state_of_mind.n.01)	1.03	0.79
	Elation (elation.n.01)	0.98	0.86
	Fugue (fugue.n.02)	0.92	0.89
	Hallucinosis (hallucinosis.n.01)	0.91	0.90

Continued on next page

Target (Synset)	Sibling (Synset)	Lin Similarity	Cosine Similarity
	Abulia (abulia.n.01)	0.85	0.76
Trauma (trauma.n.02)	Depression (depression.n.01)	1.71	0.84
	Anxiety (anxiety.n.01)	1.70	0.90
	Mental Health (mental_health.n.01)	1.55	0.86
	Mental Illness (mental_illness.n.01)	1.36	0.87
	Dissociation (dissociation.n.02)	1.33	0.84
	Hypnosis (hypnosis.n.01)	1.24	0.85
	Delusion (delusion.n.01)	1.23	0.84
	Anhedonia (anhedonia.n.01)	1.17	0.84
	Agitation (agitation.n.01)	1.15	0.90
	Depersonalization (depersonalization.n.02)	1.15	0.87
	Irritation (irritation.n.01)	1.11	0.88
	Morale (morale.n.01)	1.11	0.85
	Nervousness (nervousness.n.02)	1.10	0.85
	Enchantment (enchantment.n.02)	1.09	0.88
	Cognitive State (cognitive_state.n.01)	1.07	0.82
	State of Mind (state_of_mind.n.01)	1.07	0.85
	Elation (elation.n.01)	1.02	0.86
	Fugue (fugue.n.02)	0.95	0.89
	Hallucinosis (hallucinosis.n.01)	0.94	0.87
	Abulia (abulia.n.01)	0.88	0.82

Table 12: All Eligible Sibling Terms for Each Target Term with Lin and Cosine Similarity Scores.

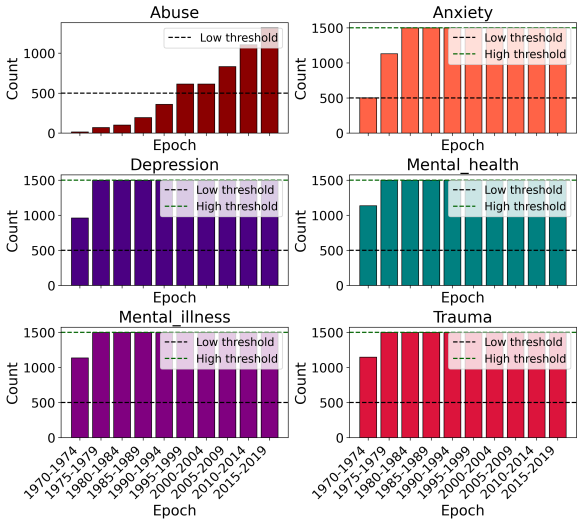

Figure 8: Counts of synthetic sentences (donor-sibling contexts).

Follow this GitHub link to access the counts of synthetic sentences for each five-year interval and the ranked lists for each sampling strategy (Boostrapped and Five-Year): [MASKED LINK]

G Multilevel Modeling Approach

To analyze the predictive effects of synthetic injections while accounting for hierarchical dependencies, we employ multilevel modeling (Gelman and Hill, 2007). These mixed linear models are well-suited for analyzing nested data, as they allow for the inclusion of fixed effects (e.g., `injection_level`) and random effects (e.g., variability across `target`). This approach leverages the full dataset while accounting for group-level structure, avoiding overfitting and unreliable estimates often encountered in simple linear regression when data points per group are limited.

Model Specifications Null Model: To assess the necessity of incorporating random effects into the analysis, we initially fit a *null model*. This null model includes only a fixed intercept (β_0) and random intercepts (u_j) to account for variability across groups (`target`), and is represented as:

$$y_{ij} = \beta_0 + u_j + \epsilon_{ij},$$

where y_{ij} is the outcome variable (e.g., `avg_valence_index_positive`) for observation i within group j, $u_j \sim N(0, \sigma_u^2)$ capturing group-level random intercepts, where $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ represents residual variability.
The intraclass correlation coefficient (ICC) is calculated to quantify the proportion of variance explained by the grouping. ICC values exceeding 0.05 indicate meaningful variability, thereby justifying the inclusion of random effects. For this dataset, the ICC is calculated as:

$$\text{ICC} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2},$$

where σ_u^2 is the variance of the random intercepts and σ_ϵ^2 is the residual variance.
Full Model: Next, we fit a *full model*, which incorporates the fixed effect of `injection_ratio` (β_1) alongside the random intercepts, expressed as:

$$y_{ij} = \beta_0 + \beta_1 \cdot \texttt{injection\_level}_{ij} + u_j + \epsilon_{ij}.$$

This full model allows us to evaluate the predictive influence of `injection_level` while accounting for hierarchical dependencies in the data.
Random Slopes Model: To further explore whether the effect of `injection_level` varied significantly across `target`, we tested an additional model with random slopes (u_{j1}) for `injection_level`, expressed as:

$$y_{ij} = \beta_0 + \beta_1 \cdot \texttt{injection\_level}_{ij} + u_j + u_{j1} \cdot \texttt{injection\_level}_{ij} + \epsilon_{ij}.$$

Here, $u_{j1} \sim N(0, \sigma_{u1}^2)$ represents the variability in slopes across groups.

Model Comparison and Selection: To determine the most appropriate model, we compare the null model, simplified random intercepts model, and random slopes model using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which provide measures of model fit, with lower values indicating better balance between fit and complexity. Likelihood ratio tests assess whether including additional random effects significantly improve model fit. Higher Log Likelihood (LL) indicates better fit.

Model Diagnostics: Residual diagnostics are performed on the final model to ensure key assumptions are met:

- **Normality:** Q-Q plots; Shapiro-Wilk test.
- **Homoscedasticity:** Residual vs. fitted value plots; Levene's test for homogeneity of variances.
- **Random Effects Variance:** Variance estimates for random intercepts and residuals to quantify group-level variability contribution.

Model results are summarized in Table 13, showing that increasing levels of synthetic injections significantly increase Dimension indices.

Method	IL	β	SE	p	LL	σ^2
Valence Index	-	-0.001	0.000	<.0001	59.85	0.02
	+	0.003	0.000	<.0001	56.30	0.03
Cosine Distance	-	NA	NA	NA	NA	NA
	+	<0.0001	<0.0001	<.0001	94.36	0.001
Arousal Index	-	-0.002	0.000	<.0001	63.45	0.01
	+	0.002	0.000	<.0001	68.38	0.009

Table 13: Results of the Final Mixed Linear Models Predicting Dimension Scores from Injection Levels

Note: IL = Injection level. β = Regression coefficient for synthetic injection level. SE = standard errors. *p*-values test the null hypothesis that the coefficient is zero. LL = Log-Likelihood, indicating model fit. σ^2 = Variance (Random Effects), quantifies variability due to grouping. NA = Not available.

Sentiment (Positive)

- The **Null Model** showed an *ICC* of 0.59, indicating that 59% of variance in `avg_valence_index_positive` is attributable to `target`, justifying its inclusion as a random intercept.

- The **Simplified Model**, with `injection_level` as a fixed effect and `target` as a random intercept, revealed a significant positive relationship ($\beta = 0.003$, $p < .0001$) and moderate variability across targets ($\sigma^2 = 0.03$). Residuals met homoscedasticity assumptions (Levene's $p = .92$), though the Shapiro-Wilk test ($p = .02$) suggested deviations from normality.
- A **Random Slopes Model**, allowing `injection_level` to vary by `target`, failed to converge, rendering the fixed effect non-significant ($p = .588$) and random slope variance negligible.
- Based on model comparison (Log-Likelihood: Simplified Model = 56.30, Random Slopes Model = 45.40; AIC/BIC unavailable due to convergence issues), the **Simplified Model** was selected as the final model (see Table 14).

Measure	Value
Number of Observations (Groups)	36 (6)
Log-Likelihood	56.30
Scale	0.0006
Random Effects Variance (Intercepts)	0.026
Fixed Effect (`injection_level`)	0.003
SE	$\pm$ 0.000
z	27.49
p-value	< 0.0001

Table 14: Model with Random Intercepts Predicting Valence Index from Positive Sentiment Injection Level.

Sentiment (Negative)

- The **Null Model** showed an *ICC* of 0.884, indicating that 88.4% of variance in `avg_valence_index_negative` is attributable to `target`, justifying its inclusion as a random intercept.
- The **Simplified Model**, with `injection_level` as a fixed effect and `target` as a random intercept, revealed a significant but minimal negative relationship ($\beta = -0.001$, $p < .0001$) with notable variability across targets ($\sigma^2 = 0.021$). Residuals met homoscedasticity assumptions (Levene's $p = .930$), while the Shapiro-Wilk test ($p = .039$) suggested minor deviations from normality, confirmed negligible by Q-Q plots.
- A **Random Slopes Model**, allowing `injection_level` to vary by `target`, failed to converge due to overparameterization or insufficient data. The fixed effect became non-significant ($p = .822$), random slope variance was negligible ($\sigma^2 = 0.006$), and the covariance between group-level variability and `injection_level` was near zero (-0.000), indicating no significant interactions.
- Given model comparison results (LL: Simplified Model = 59.85, Random Slopes Model = 48.90; AIC/BIC unavailable due to convergence issues), the **Simplified Model** was selected as the final model (see Table 15).

Measure	Value
Number of Observations (Groups)	36 (6)
Log-Likelihood	59.85
Scale	0.0005
Random Effects Variance (Intercepts)	0.021
Fixed Effect (`injection_level`)	-0.001
SE	$\pm$ 0.000
z	-11.40
p-value	< 0.0001

Table 15: Model with Random Intercepts Predicting Valence Index from Negative Sentiment Injection Level.

Intensity (High)

- The analyses assess the impact of `injection_level` on `avg_arousal_index_high`. The Null Model (*ICC* = 0.54) justified including `target` as a random intercept, as 54% of variance was attributable to group differences.
- The **Simplified Model with Random Intercepts**, incorporating `injection_level` as a fixed effect, converged and showed a small but significant positive effect ($\beta = 0.002$, *SE* = 0.000, $p < .0001$), with notable group-level variability ($\sigma^2 = 0.009$).
- A **Random Slopes Model**, allowing `injection_level` to vary across groups, failed to converge due to overparameterization or insufficient data. Its exceptionally low scale (0.0002) suggested overfitting or model specification issues.
- Based on convergence and parsimony, the **Simplified Model with Random Intercepts** was selected as the final model. LL metrics

(Simplified: 68.38, Random Slopes: 59.99) confirmed this choice (see Table 16).

Measure	Value
Number of Observations (Groups)	36 (6)
Log-Likelihood	68.38
Scale	0.0003
Random Effects Variance (Intercepts)	0.009
Fixed Effect (`injection_ratio`)	0.002
SE	$\pm$ 0.000
z	23.63
p-value	<.0001

Table 16: Model with Random Intercepts Predicting Arousal Index from High Intensity Injection Level.

Intensity (Low)

- The Null Model (*ICC* = 0.53) justified including `target` as a random intercept, as 53% of variance was attributable to group differences.
- The **Simplified Model with Random Intercepts**, incorporating `injection_level` as a fixed effect, converged and showed a small but significant negative effect on `avg_arousal_index_low` ($\beta = -0.002$, *SE* = 0.000, $p < .0001$). Random intercept variance ($\sigma^2 = 0.011$) indicated notable group-level variability. Model assumptions were met: homoscedasticity (Levene's test, $p = .982$) and linearity, though residuals deviated from normality (Shapiro-Wilk, $p = .002$), with minimal impact on Type I error given large effect sizes.
- A **Random Slopes Model**, allowing `injection_level` to vary across groups, failed to converge due to overparameterization or insufficient data. Its exceptionally low scale (0.0002) suggested overfitting or model specification issues.
- The **Simplified Model with Random Intercepts** was selected as the final model, supported by LL metrics (Simplified: 63.45, Random Slopes: 58.63). See Table 17.

Measure	Value
Number of Observations (Groups)	36 (6)
Log-Likelihood	63.45
Scale	0.0004
Random Effects Variance (Intercepts)	0.011
Fixed Effect (`injection_level`)	-0.002
SE	$\pm$ 0.000
z	-22.98
p-value	<.0001

Table 17: Model with Random Intercepts Predicting Arousal Index from Low Intensity Injection Level.

Breadth

- The analyses assess the impact of `injection_level` on `cosine_distance_mean`. The **Null Model** (*ICC* = 0.71) justified including `target` as a random intercept, as 71% of variance was attributable to group differences.
- The **Simplified Model with Random Intercepts**, incorporating `injection_level` as a fixed effect, converged and showed a small but significant positive effect ($\beta < 0.0001$, *SE* < 0.0001, $p < .0001$), with notable group-level variability ($\sigma^2 = 0.001$).
- A **Random Slopes Model**, allowing `injection_level` to vary across groups, failed to converge, likely due to overparameterization or insufficient data. The model's scale was near zero (<0.0001), suggesting overfitting or misspecification.
- Based on convergence and parsimony, the **Simplified Model with Random Intercepts** was selected as the final model. LL metrics (Simplified: 94.36, Random Slopes: 84.01) confirmed this choice (see Table 18).

Measure	Value
Number of Observations (Groups)	36 (6)
Log-Likelihood	94.36
Scale	0.0001
Random Effects Variance (Intercepts)	0.001
Fixed Effect (`injection_level`)	<0.001
SE	$\pm$ <0.0001
z	7.49
p-value	<.0001

Table 18: Model with Random Intercepts Predicting Arousal Index from High Intensity Injection Level.

H SIB Scores: Results for Five-Year Random Sampling Strategy

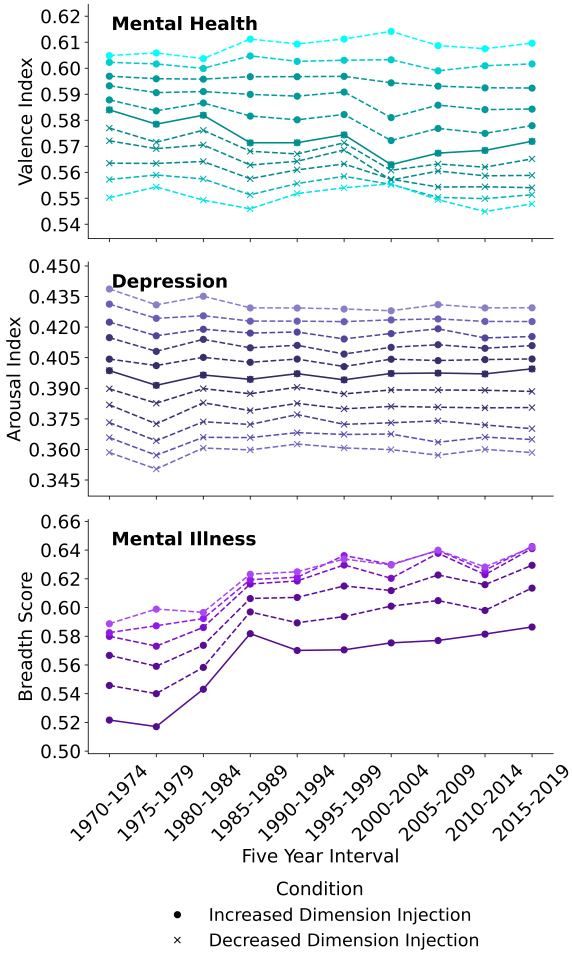

Figure 9: SIB Scores by Five-Year Intervals Across Injection Levels and Conditions.

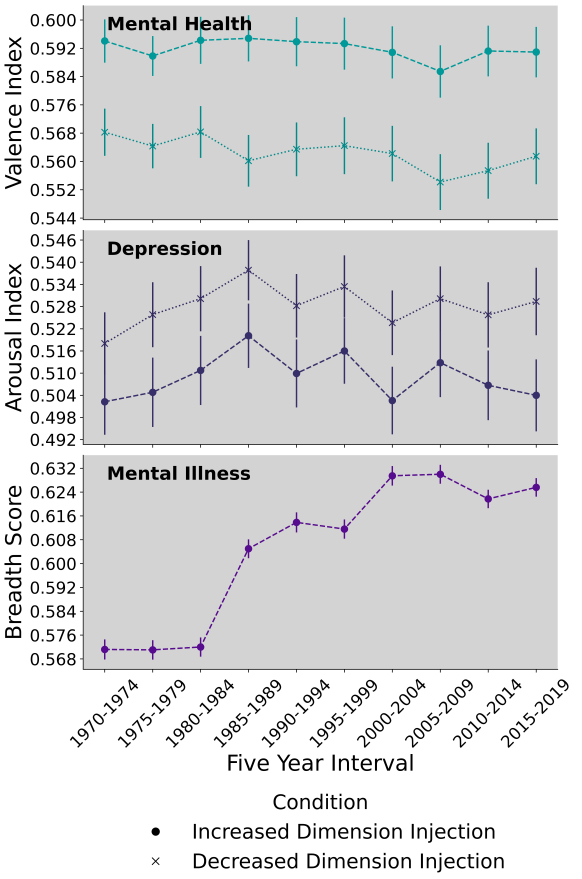

Figure 10: SIB Scores (±SE) by 50% Injection Levels and Conditions: Control Setting for Five-Year Samples.

I Alternative LSC Detection Methods: Results for Bootstrapped Settings

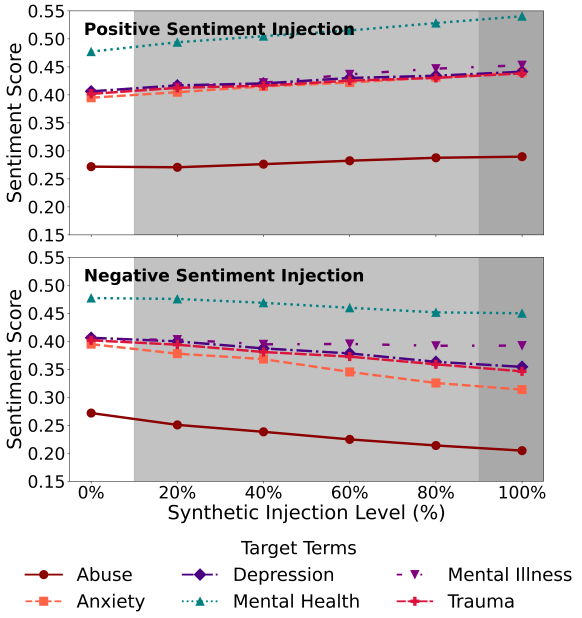

Figure 11: ABSA Sentiment Index Across Injection Levels and Sentiment Conditions: Bootstrapped Samples.

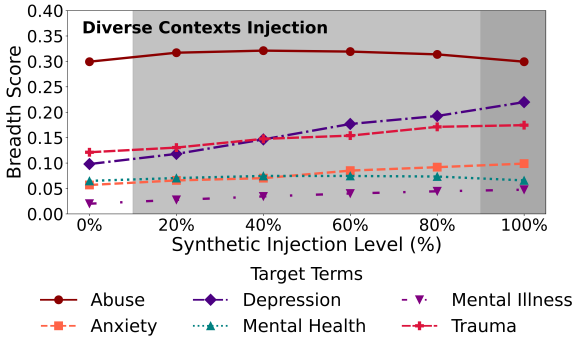

Figure 12: XL-LEXEME Breadth Score (Average Cosine Distance Within-Bins) Across Injection Levels and Breadth Condition: Bootstrapped Samples.

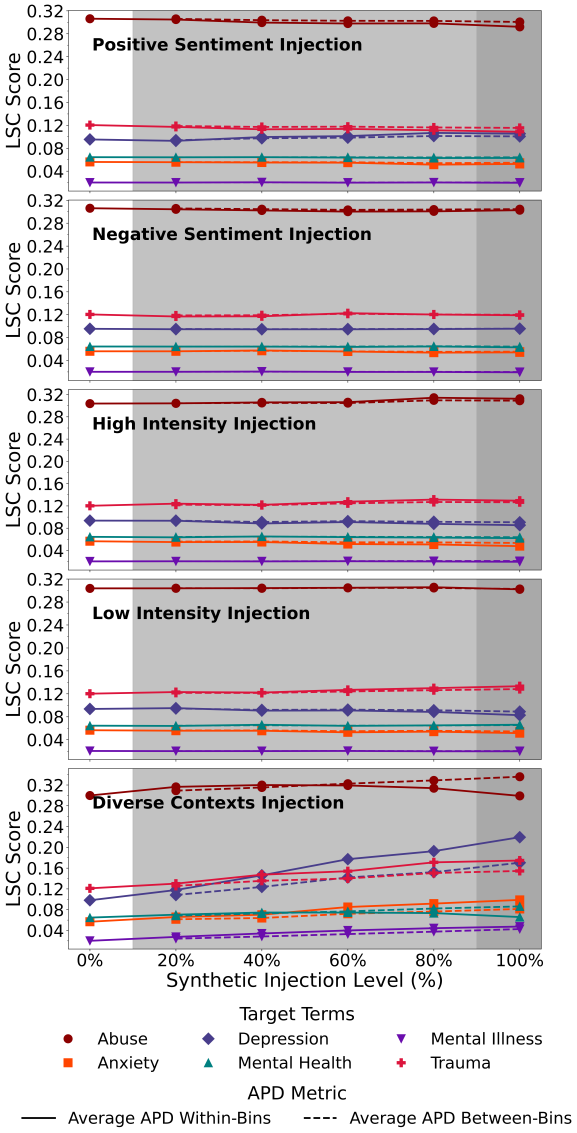

Figure 13: LSC Scores (APD Between-Bins and APD Within-Bins) Across Injection Levels and SIB Conditions: Bootstrapped Samples.